

Time for a bipartisan OTA

The US legislature is bereft of objective guidance on issues that underpin much of its work. A congressional Office of Technology Assessment should be reinstated as soon as possible, on a solid basis of bipartisan support.

When the US Congress abolished its Office of Technology Assessment (OTA) back in 1995, there was much talk of the most powerful legislature in the world obtaining advice on technical and scientific issues from “a multiplicity of sources” elsewhere. These sources, it was claimed, would provide the variety of perspectives necessary for Congress to deal with the complex science- and technology-based issues that fall under its jurisdiction.

Unsurprisingly, perhaps, nothing akin to this has materialized. The OTA has left a vacuum, as its defenders said it would. In the six years since it disappeared, difficult science-based issues continue to confront the legislature — global warming, transgenic crops, the alleged energy crisis and embryonic stem-cell research, to name but a few. The OTA was meant to provide an objective assessment of these kinds of issues. What Congress receives instead is a deluge of information from outside sources, mostly biased, some of it masquerading as independent.

The OTA was established in 1973 to help harassed congressmen and -women — and, yes, many of them do work 18-hour days, whatever it says in the movies — and their non-specialist staff to digest information on issues that are inherently complex. Unfortunately for this ideal, the OTA operated during a period in which the Democrats controlled the House all of the time and the Senate almost all of the time. Frustratingly for the OTA's ambitions to establish a non-partisan reputation, it was acting as a counterweight to a Republican administration in the White House for all but four of these years.

It is no wonder that at the end of that time, the agency had become

associated with a Democrat agenda in the minds of some Republicans. Additionally, when the Republicans took control of Congress in 1995 with a radical government-cutting agenda, they soon discovered that it was almost impossible to close down any component of the US government, and picked on their own scientific brains trust as just about the only thing they could shut. It was a foolish decision, resisted at the time by some Republicans who knew the OTA well, such as Amo Houghton (Republican, New York), and regretted now by others.

A modest effort is under way to prepare the groundwork for the revival of something similar to the OTA. Academics and former staff of the office will meet in Washington on 14 June to discuss a strategy for this.

But great care will be required if anything is to come of this effort. The most important challenge is to win the confidence of a substantial body of Republicans in any proposed model. There are idealogues in Congress who will have no appetite for an instrument such as the OTA, however objective and efficient it may be. But they are not in the majority in either body of Congress, and their hostility should not discourage steps to revive the agency.

On the contrary, a technology-assessment office should be revived now, on a truly bipartisan basis. Allies of the concept may be tempted to bide their time, aware that there is a good chance that the Democrats will regain control of the House and Senate next year. Such a delay would be a lost opportunity. A small, dynamic and politically neutral office established now — while Republicans control Congress — would stand a better chance of avoiding the fate of the last one. ■

An end to procrastination?

A new German bioethics council should lead to a prompt resolution of debates over stem cells.

Germany's main grant-giving agency, the Deutsche Forschungsgemeinschaft (DFG), is proposing a much-needed relaxation of the country's restrictive embryo protection law, which bans the cultivation of human stem-cell lines from embryo cells. In doing so, it crosses swords with a hostile government whose position on the sanctity of embryonic cells appears unshakable (see page 119).

Coincidentally, Chancellor Gerhard Schröder has created a national ethics council to advise the government on bioethics (see page 124). Stem-cell research will be its baptism of fire, and could be a precedent for how Germany handles future ethical dilemmas in basic research.

The DFG has not rushed its thinking on human embryonic stem-cell research. It has taken due time to consider the issue thoroughly, and has reached fair and balanced conclusions. Its suggestion that, in well-justified cases and with appropriate checks and controls, researchers should be allowed to import and even to cultivate human embryonic stem-cell lines is a reasonable response to recent scientific advances. These advances increasingly indicate the medical potential of stem cells to replace diseased tissues and organs.

It is true, of course, that the DFG's position is driven by the interests of the scientific community. But this does not mean that such interests are inevitably different from those of society, which will probably not want to miss out on any medical benefits that might arise.

Appropriately and inevitably, German citizens want any move towards the use of humans — including the helpless embryo — for research purposes to be discussed in depth. Soul-searching debate on embryonic stem-cell research has filled newspapers and spawned a stream of bioethical conferences over the past two years. The DFG has held back from endorsing the research, arguing for its restriction to adult stem cells until the weight of scientific information justified crossing a new boundary — as it now believes is the case.

But when does society's wider debate end? The government's demand for it to continue smacks of fear of taking a tough decision the year before a general election. Hopefully, the new national ethics council will not feel the need to start from scratch, or unduly delay reporting its conclusions. As Germany learnt to its cost in gene sequencing, late entry to a fast-moving research field leaves the research community and others at a significant disadvantage. ■





Star wars
Physicists non-plussed as Bush unveils defence plans
p120



Say cheese
Public consortium produces draft mouse sequence
p121



Nerve centre
Britain launches investigation into wartime tests
p122



Primate lab
Europe ends experiments on chimpanzees
p123

Stem-cell research in doubt as funders clash with government

Alison Abbott, Munich

A public clash between Germany's main research funding agency and the government over human embryonic stem-cell research has stalled a decision on the country's first grant application for such work.

The agency, the Deutsche Forschungsgemeinschaft (DFG), announced on 3 May that it was ready to fund human embryonic stem-cell research. But the research ministry, the BMBF, responded immediately with a call for a moratorium on funding individual projects while ethical and moral issues are clarified — setting the stage for a rare confrontation between the supposedly independent DFG and the BMBF, which provides about half of its funding.

As each side paused to consider its position, the DFG's grants committee (which includes BMBF representatives) deferred a decision on the grant to the University of Bonn neuroscientist Oliver Brüstle.

A German embryo protection law forbids creation of stem-cell lines from human embryos for research purposes. But it is not



Taking a stand: the DFG's Ernst-Ludwig Winnacker (right) sees no reason why work on imported embryonic stem cells should not receive funding.



illegal to import cell lines for such purposes. Two years ago the DFG published a position statement saying that, given the special sensitivities of German society and how little was known about how embryonic and adult stem cells differentiate into other cell types, German scientists should focus their efforts on

studying the potential of adult stem cells.

But research has moved much further since then, says Ernst-Ludwig Winnacker, president of the DFG — “and surprisingly fast”. After its senate meeting last Thursday, the DFG issued a new position statement, saying that it now saw “no justification for excluding [from funding] research on imported embryonic stem cells produced legally in other countries.” The statement also hints at the need for the embryo protection law to be modified to allow cell lines to be developed in Germany.

The DFG's grants committee, which comprises 19 academic members and 18 representatives from DFG funding organizations, primarily the federal and state governments, met on Friday to discuss funding of applications ranked through peer review. It had been expected to approve Brüstle's application, submitted 10 months ago, to study *in vitro* differentiation of human embryonic stem cells into neuronal cells, for transplantation into myelin-deficient rats. Deficiency in myelin, the sheath that protects nerves, occurs in multiple sclerosis.

But just after the DFG had announced its new position, research minister Edelgard Bulmahn said that the BMBF “would apply pressure” to defer Brüstle's application.

Winnacker denies experiencing political pressure. “The committee had had no time

Reprieve for Smithsonian centre

Corie Lok, Washington

The Smithsonian Institution has abandoned plans to close the National Zoo's Conservation and Research Center in Front Royal, Virginia (see *Nature* 410, 727; 2001), after an outcry from conservation biologists and local politicians.

Smithsonian secretary Larry Small, announcing the reversal on 6 May, said the closure plan had been “misinterpreted” by the public as an abandonment of science. He claimed that keeping the centre open was “necessary to correct that false perception”.

Small said that the Board of Regents of the museum complex had approved his other plans to restructure museum research and focus it on “centres of excellence”. But he announced that a panel of scientists would be set up to advise him on the details of the restructuring. Its members will be drawn

from inside and outside the institution.

A subsequent regents' meeting, chaired by former senator Howard Baker, approved Small's reorganization recommendations in principle and the creation of the panel.

Brian Huber, a palaeobiologist at the National Museum of Natural History and chair of the Smithsonian's academic senate, says Smithsonian researchers “feel pretty targeted” by the restructuring proposal. He wants the panel to look at the whole museum complex, not just at science. “Let's not focus cuts and restructuring just on research,” says Huber.

The regents' meeting did approve the closure of the Smithsonian Center for Materials Research and Education, together with the video and audio production facility and two offices at the museum libraries, with the loss of 180 jobs.

to absorb the conclusions of the senate the night before, and we decided it would only be fair to give the evaluators time to consider the implications," he says. The decision was deferred for two months.

Bulmahn says that Germany's new national ethics council, created last week by Chancellor Gerhard Schröder, should debate the issue of human embryonic stem-cell research. "New courses that cross long-established ethical boundaries, cannot be changed in a hurry," she says.

Wolf-Michael Catenhusen, state secretary for research, says: "The DFG can formulate its position as often as it wants, but this does not change public opinion — nor political opinion."

"It is very appropriate that Germany has approached embryonic stem-cell research slowly and carefully," says Brüstle. But he is surprised and disappointed that his research is still being held up. "The continual delays are starting to feel no different from rejection."

Winnacker declines to speculate on what would happen if the research ministry tried to instruct the DFG — whose independence is enshrined in statute — not to fund embryonic stem-cell research. "I don't know," he says. "Political interference is something that has never happened in the 51 years of DFG history, so we have no precedent." ■

Birthday meeting bemoans low profile of science office

Steve Nadis, Boston

It may be 25 years old, but the White House Office of Science and Technology Policy (OSTP) has yet to establish a strong identity in Washington. That was one recurrent theme at a conference held on 1 May to celebrate the OSTP's anniversary.

John Porter, a science advocate and former Illinois congressman, told the meeting that most members of Congress "would be unlikely to have ever heard of the Office of Science and Technology Policy or to know the initials OSTP. They would probably recognize that the president has a science adviser, but few would be able to identify him or her."

Held at the Massachusetts Institute of Technology, the meeting attracted seven former presidential science advisers, including 91-year-old William Golden, who served as the first real science adviser to Harry Truman just after the Second World War. The position persisted in various forms before it was institutionalized in the 1976 law that established the OSTP as a permanent office, the director of which also serves as the president's science adviser.

Porter said that the low visibility of the OSTP placed an even stronger onus on scientists to step forward and make themselves



John Porter says scientists need to be heard.

heard in policy debates. MIT president Charles Vest agreed, calling for a new generation of "citizen scientists to explain what we do and why it is important". Changes in education are needed to prepare scientists for this additional role, he said.

Some of the meeting's organizers had hoped that the new administration of President George W. Bush would appoint a science adviser in time for the event, but it did not. Porter noted that Bush has already made decisions on matters such as the Kyoto Protocol and missile defence without input from a presidential science adviser or from the OSTP.

The lack of a science adviser at this stage of the administration is not unprecedented, meeting attendees were told, but it still concerned some of the speakers. Golden said it mattered "because science and technology runs through the fabric of so many issues on both domestic and foreign fronts. Functioning without a science adviser," he said, "is like playing baseball without a full team." ■

Plans for missile defence system perturbs physicists

Irwin Goodwin, Washington

The announcement by President George W. Bush that he plans to abandon the 1972 Anti-Ballistic Missile (ABM) treaty has been met with concern from leading US physicists.

The reasoning behind Bush's plan to move ahead rapidly with a new missile defence system for the United States and its allies was questioned by physicists at a recent meeting of the American Physical

Society (APS) in Washington.

"The last thing the so-called rogue countries would do is launch against the United States, because of the instant retaliation that would result," Richard Garwin, a retired IBM physicist who has advised the US government on ABM and other technical issues, told the meeting.

Bush has so far offered few details about what his missile defence structure would look like, when it will be deployed or what it would cost. But he has said that the system would be "layered", meaning that it would involve attempts to intercept missiles at each of their flight stages — during launch, in orbit and during re-entry.

Garwin says that the best prospect of a working system is one that targets the missile during its boost phase, as it moves into orbit. The APS has set up a 12-member panel to look into the feasibility of this approach and deliver an unclassified report by the end of the year, in time to feed into Bush's decision on what systems to build.

The Clinton administration's far less ambitious plan called for a smaller system,

with 100 missile interceptors based in Alaska and intended to protect against a limited attack by a 'rogue state'. The Congressional Budget Office estimated that Clinton's system would cost about \$60 billion to deploy.

Some analysts say that the United States has already spent about \$60 billion on missile defence over the past 50 years, including \$27 billion on President Ronald Reagan's Strategic Defense Initiative, with little to show in return.

Spending on missile defence this year is expected to top \$4.9 billion, making it the largest single programme in the defence department, even before the programme is expanded to meet Bush's objectives. Critics say that the Bush initiative is likely to cost \$100 billion or more.

Beneficiaries of the programme may include university scientists in fields such as optics and nanotechnology, but the great bulk of the expenditure will be in development, rather than scientific research, according to officials in the defence department. ■



Bush reveals his plans for missile defence.

Dissent grows over Helmholtz proposals

Alison Abbott, Munich

Scientists at the German Cancer Research Centre (DKFZ) in Heidelberg are stepping up their protests against proposed reforms at the Helmholtz Society, to which the centre belongs. The reforms are supposed to usher in a more competitive research regime.

On 26 April, top researchers at the centre, one of the 16 government-funded national research laboratories that make up the Helmholtz Society, wrote an open letter of protest to the research minister Edelgard Bulmahn, claiming that the reforms threatened their academic freedom.

The changes were suggested last year by the government as a way to restore productivity and industrial relevance at the Helmholtz laboratories. But the DKFZ researchers say the reforms will instead bind the labs in red tape and submit them to the control of management boards dominated by non-scientists.

They complain that their concerns over how the Helmholtz labs will be funded have been ignored. The scientists first put their objections to the DKFZ board and to heads of other Helmholtz centres last November.



Basic problems: Werner Franke fears a future of red tape if the Helmholtz Society reforms proceed.

Under the reform plan, an Association of Research Centres of the Helmholtz Society would be set up to design a coordinated research programme for the centres. Funding would be distributed by the federal research ministry according to six strategic areas of research — health, energy, environment, transport and space, structure of matter and ‘enabling’ technologies (such as new

research tools). Research programmes would be evaluated every five years, with regular milestones being set.

Werner Franke, head of the DKFZ’s Division for Cell Biology, says such centralized planning is reminiscent of the way the East German Academy of Sciences used to operate. “We see a breathtaking level of really Germanic red tape being introduced,” he says.

Franke also thinks the proposed association lacks the right mix of expertise. It will consist of the research minister, two research ministers from state governments, six external scientists and six representatives of industry as well as representatives from other German research organizations and from the finance ministries. “There are too many industrialists who do not understand the needs of basic research,” he says.

But Wolf-Michael Catenhusen, state secretary for research, says the changes will bring the centres in line with other countries. “There are really no grounds for bringing in a 1970s-style ideology debate to something that is simply introducing a modern instrument for administering research,” he says.

Detlev Ganten, president of the Helmholtz Society, says only a few scientists believe their academic freedom is under threat. Moreover, the research centres have agreed to join the new association on condition that each centre is allowed flexibility in its budget. This provision is expected to be in place this year.

Rudi Balling, director of the German Centre for Biotechnology in Braunschweig, says that the centres’ acceptance of the plan is absolutely contingent on this. “We need to be able to shift assigned budgets between projects, if the scientific results so dictate,” he says.

Martin Lipp, an immunologist who chairs the scientific advisory council at another Helmholtz centre, the Max Delbrück Center for Molecular Medicine in Berlin, says that: “Overall, if carried out optimally, the restructuring is a positive thing which will enable us to respond better to developments in science.”

Mouse genome effort ‘on course’

Jonathan Knight, San Francisco

Public genome sequencers say they have met the goal they set for themselves last October, by producing crude sequencing data for almost the entire mouse genome.

But they haven’t yet figured out how to finish the project to the level of accuracy they would like to see, which would mean sequencing each base in the genome an average of 10 times over.

The Mouse Sequencing Consortium has been using the high-throughput ‘shotgun’ method pioneered by the private company Celera Genomics of Rockville, Maryland. This technique involves sequencing some 6 million fragments, each between 500 and 700 bases long, of the mouse genome and then assembling them later.

The public consortium says it has now sequenced 94% of the genome, hitting each base an average of three times. But the sequence data remain almost completely unassembled. The public database contains more than 15 million unordered fragments.

The consortium’s announcement came just 10 days after Celera revealed that it had completed and assembled 2.6 billion bases of the mouse genome (*Nature* 411, 8; 2001). Unlike the Celera data, the public data are available free to all researchers.

But the next phase of the project remains undefined. This was the plan all along, says

Jane Peterson, a project official at the National Human Genome Research Institute (NHGRI), because the partial sequence data will be useful in working out which techniques to use to complete the genome.

Three corporate sponsors paid a quarter of the \$58 million cost of the initial phase. But so far no arrangements have been made for continued private support of the project. And although the NHGRI has ample funds to continue sequencing the mouse, no budgets for the next phase have yet been set.

“At this point we are shifting strategies,” says Robert Waterston, director of the Washington University genome sequencing centre. “We are developing plans to get the complete genome sequence.”



Rough rodent: the public consortium has announced an early draft of the mouse genome.

Canada stakes claim on fusion energy project

Rex Dalton, San Diego

Canada's bid to construct the Iter fusion energy research facility is building momentum, with moves to obtain governmental permits and support for the international collaborative project.

Iter Canada, a consortium that wants to build the multibillion dollar facility to study fusion as an energy source, started the environmental review process this week for a site at Clarington, near Toronto.

Backers of the project hope that Canadian prime minister Jean Chrétien may soon endorse Canada's project, which is competing against sites in Japan and France (see *Nature* 410, 856; 2001).

Iter, which originally stood for International Thermonuclear Experimental Reactor, is a joint effort between Japan, Europe and Russia to create an energy source by producing nuclear fusion in plasma confined magnetically in a doughnut-shaped chamber known as a tokamak.

The United States withdrew from the project in 1998, after Congress made large reductions in its fusion energy research programme. Canada's bid is seen as strategically significant, as it could potentially entice the United States back into the project. The costs would be borne by all the international participants.

Efforts are still under way in the United

States to boost fusion research funding. Representative Zoe Lofgren (Democrat, California), for example, is planning to introduce a bill that would revive US participation in Iter and inject a massive \$320 million into the project next year.

The Iter facility would conduct research intended to prove, in principle, that the tokamak could serve as a viable energy source. This device would burn plasma for a set period of time, although its specification has been cut back from an earlier one that was intended to ignite the plasma by making its burning self-sustained.

Formed three years ago, Iter Canada has

been garnering support for the fusion facility among Canadian businesses, labour groups and universities. The consortium recently incorporated the non-profit-making International Fusion Energy Institute, the body that filed for an environmental assessment of the fusion project by the Canadian Nuclear Safety Commission. This week, the commission's staff began preparing an environmental assessment of the project.

Iter Canada's chairman, Peter Barnard, noted that a host site must demonstrate that it can be licensed. "We decided the best way to proceed is to get a licence," says Barnard, an engineer based in Toronto. After securing the appropriate environmental permit, Iter Canada faces a further government licensing process, which could take several years.

Iter Canada's proponents say support for their site is growing over the French and Japanese locations. Building the facility in Canada will be cheaper, they say, because the site is so well prepared and construction costs in the country are relatively low, leaving more funds for research.

The site at Clarington is near four nuclear reactors that produce tritium, Iter's fuel, as a by-product. Tritium is also an important component of nuclear weapons, and the requirement to handle it is seen as a potential impediment to Japan's bid to host Iter.

http://itercanada.com



Inquiry set up into Porton Down nerve-gas death

David Adam, London

The British Ministry of Defence (MOD) is launching a new investigation into the circumstances surrounding the death of an airman who took part in military experiments with chemical warfare agents nearly 50 years ago.



Maddison: a victim of sarin?

Ronald Maddison died in 1953, allegedly after the nerve agent sarin was dripped onto a patch of uniform taped to his arm in experiments aimed at establishing toxicity thresholds.

The MOD said on 1 May that it will carry out a 'historical survey' of its service volunteer programme, which ran from 1939 to 1989 at

Porton Down in Wiltshire. Formerly the site of controversial research on chemical and biological weapons, the centre remains active, working on defences against such weapons.

There have been claims that some volunteers were duped into taking part in nerve-gas trials, thinking they were helping to find a cure for the common cold. About 250 veterans who were involved in tests at Porton Down have formed a support group and have called on the MOD to investigate the possible health effects of the trials.

The review will be supervised by Ian Kennedy, a respected medical ethicist at University College London. An MOD spokesperson says it will cover "absolutely everything surrounding the programme — who took part, how the volunteers were recruited, the way in which the programme was described to them and what they were told was taking place".

The announcement of the MOD review comes two months after the coroner, the local official responsible for enquiries into violent or accidental deaths, rejected the 1953 misadventure verdict returned on Maddison's death by an inquest held behind closed doors. That decision followed a two-year police inquiry into events at Porton Down. The coroner's office is now awaiting a

decision from the Attorney-General in London on a request to reopen the inquest, but does not expect a reply this year.

The MOD says its review is "not particularly related" to the coroner's desire to reopen the inquest, but that it is responding to public interest generated by the police inquiry. The reviewers will report next spring to the MOD, which says it will publish the findings.



In defence: Porton Down, previously the site of biological and chemical weapons research.



The chimps are down: macaques will take centre-stage at the Biomedical Primate Research Centre.

Europe brings experiments on chimpanzees to an end

Sally Goodman, Paris

The only research facility in Europe conducting experiments on chimpanzees is to end all research using the apes.

The decision to stop the experiments at the Biomedical Primate Research Centre (BPRC) in Rijswijk, the Netherlands, comes after years of dissatisfaction over funding, and criticism from animal-rights groups over the chimps' living conditions.

A report commissioned for the Dutch science ministry, made public on 27 April, said that research on chimpanzees housed at the centre should be phased out and the animals rehoused.

Anton Berns, head of research at the Netherlands Cancer Institute and chairman of the committee that produced the report, says: "Researchers at the BPRC are also victims of the poor facilities. They have been unable to improve the situation because of the lack of funding."

The centre, which houses around 1,300 non-human primates including 108 chimpanzees, is currently funded by the Dutch science ministry to the tune of 5.2 million guilders (US\$2.1 million) per year—about a quarter of its running costs. The rest comes from the European Commission and from private research contracts. Much of the centre's research is on vaccines for AIDS and hepatitis C, although malaria and other parasitic diseases are also investigated.

Some researchers have attacked the decision. One British scientist, who declined to be identified, said: "For years, we were positively encouraged by the European Commission to use the BPRC as a core research facility. We will be losing a valuable resource."

BPRC's director Ronald Bontrop says the unit will accept the committee's recommendation to end chimp research, provided that the income needed to assure its future as a "quality research centre", working with other primates, can be guaranteed.

The committee accepts that work with chimps on hepatitis C is justified as they are the only other species to carry the virus—but it suggests that the work be done with larger chimp colonies in the United States. It recommends that hepatitis work at Rijswijk be phased out over two or three years, but that AIDS work ceases immediately.

The future of the chimps is uncertain. The committee says that those infected with viruses should be given special housing, and healthy animals should go to sanctuaries or zoos, using funding from the Dutch government and the European Commission.

Research on the other primates at the centre, notably rhesus macaque monkeys, will continue. But the committee says that more of the work should be conducted in peer-reviewed academic projects rather than under private contracts. It also recommends that the centre, which is currently managed independently, should establish strong links with a university. Leiden or Rotterdam universities are the most likely candidates, says Bontrop.

Hundreds of chimps have been bred in captivity around the world in the past 20 years to provide models for health research, mainly into AIDS vaccines. But it is now generally accepted that rhesus macaques provide a better model for HIV.

This leaves primate centres trying to cope with unneeded chimps, which are expensive to maintain and the culling of which would be unacceptable to the public. And some AIDS researchers still believe that chimps provide a uniquely useful animal model. Understanding why chimps stay healthy when infected with SIV, the simian equivalent of HIV, while humans die of AIDS is "a question of enormous importance", says Mark Feinberg, an immunologist at Emory University in Atlanta, Georgia. "It needs to be answered whether the research is performed in the Netherlands or not," he says. ■

Record donation from Silicon Valley pioneer goes to Stanford

Jonathan Knight, San Francisco

Stanford University has received a gift of \$400 million from the foundation established by Hewlett-Packard founder William Hewlett. The donation is thought to be the largest ever made to an institution of higher education.

The William and Flora Hewlett Foundation, which announced the gift on 2 May, hopes it will inspire further donations to universities.

Hewlett was a Stanford graduate who in 1936 co-founded Hewlett-Packard, the electronics company that was to play a central role in the development of Silicon Valley. He died in January at the age of 87, but his son Walter, who is chairman of the Hewlett Foundation, said the gift had been planned for over a year.

A quarter of the money is earmarked for undergraduate education. The rest will go to the School of Humanities and Sciences, the largest of Stanford's seven schools but one less well-endowed than others such as medicine or engineering.

The School of Humanities and Sciences includes most of the university's basic science departments and research labs. Walter Hewlett, who also serves on the school's advisory council, said that it was threatened by budgetary pressures.

"It became clear to me that if something wasn't done to put the school on a more firm financial footing, there would be a major deterioration," he said.

William Hewlett and Hewlett-Packard co-founder David Packard, also a Stanford graduate, had already given some \$300 million to the university over the years.

According to the *Chronicle of Higher Education Almanac*, the only larger gift to higher education on record is the \$1 billion pledged by Microsoft chairman Bill Gates in 1999 for the Gates Millennium Scholars Program. ■



Sunny outlook: William Hewlett's donation will aid humanities and science at Stanford.

Bone-marrow cells complicate stem-cell funding debate

Washington In a discovery that is likely to be seized upon by opponents of embryonic stem-cell research, biologists in the United States have identified bone-marrow cells that can develop into a wide range of tissues, including skin and the cells that line guts and lungs. The discovery by researchers led by Diane Krause of Yale University and Neil Theise of New York University (*Cell* **105**, 369–377; 2001) demonstrates the broad developmental potential of ‘adult’ stem cells.

Opponents of embryo research argue that biologists trying to develop stem-cell therapies do not need to work on human embryonic stem cells, which can be isolated only by destroying a pre-implantation embryo. Most researchers think that work on both embryonic and adult cells is needed, but fear the US administration will deny funding for embryonic stem-cell research.

Germany sets up national bioethics council

Munich Germany last week set up a council on bioethics that will report directly to Chancellor Gerhard Schröder. Top of its

agenda are pre-implantation genetic diagnostics, stem-cell research and embryo protection. The council’s 23 members — scientists, sociologists, lawyers, theologians, trade unionists and representatives of industry and the employers’ federation — are appointed for four-year terms.

“A national ethics council has been overdue,” Wolf-Michael Catenhusen, state secretary for research, told *Nature*. “It will hopefully help us catch up with the United States and France, where similar commissions have successfully triggered very fruitful public debates.”

Harvard to look into alternative medicine

Boston Complementary medicine is to come under fresh scientific scrutiny, thanks to a \$10-million grant to Harvard Medical School. The grant, from the Bernard Osher Foundation, a San Francisco-based philanthropic organization, will be used to create a new research centre to be headed by Harvard associate professor David Eisenberg. The centre will assess the safety, efficacy, cost-effectiveness and mechanisms of action of acupuncture, chiropractic, herbal medicine, and other remedies.

In 1998 the Osher foundation established a similar centre at the University of California,



No alternative: a new centre at Harvard will examine such treatments as acupuncture.

San Francisco. Eisenberg hopes that, as more universities investigate unconventional therapies, “we will forget the terms ‘alternative’ and ‘complementary’ altogether and simply provide the best available medicine, based on the best available information”.

Biomedical scientist wins top French research post

Paris France has a new research supremo, in the form of Ketty Schwartz of the Pitié-Salpêtrière Hospital in Paris, best known for her work on repairing damaged heart tissue using grafts of muscle cells. Schwartz’s appointment as director of research in the research ministry confirms last year’s

statements from the minister, Roger-Gérard Schwartzberg, that the biological sciences would be his top priority. It also demonstrates his commitment to appoint more women to top positions.

Schwartz replaces Vincent Courtillot, who was appointed by the previous minister, Claude Allègre. Both were Earth scientists. Schwartzberg is not a scientist, so Schwartz is expected to have a key role in developing French science policy.

Three million log on in hunt for aliens

Washington SETI@home, the project that enables ordinary computer users to join forces to analyse signals received by radiotelescopes in search of extraterrestrial intelligence, expects to attract its three millionth user this week.

The project, based at the University of California at Berkeley, has chalked up 650,000 years of computer-processing time since it began in May 1999 — making it by far the largest computational task ever performed.

There's no sign yet of any extraterrestrial radio channels among the intergalactic noise. But the project — the first to hit on the idea that idle personal computers could be combined into truly massive parallel

supercomputers — has spawned plenty of imitators. The Intel-United Devices Cancer Research Project, for example, which screens molecules for those most likely to interact with proteins that are involved in cancer, has attracted more a third of a million members since it started earlier this year.

► <http://setiathome.berkeley.edu>

► <http://www.ud.com>

Europe launches virtual maritime institute

Munich European research commissioner Philippe Busquin's much-touted plans to create a European Research Area have taken a step forward with the launch of the first 'virtual institute' created under the scheme.

The European Virtual Maritime Institute (EVIMAR) brings together 17 partners — both companies and research institutes scattered across the continent — to exchange information about ongoing projects, scientific developments and research needs. Its primary goal is to enhance the competitiveness of Europe's marine technology and shipbuilding industries.

Pierre Papon, former head of IFREMER, the French national maritime research agency, applauds its creation, but argues that further European collaboration is needed. "EVIMAR is a modest positive step," he says.

"But at the end of the day Europe needs a strong central maritime science and technology agency."

Lost NASA spacecraft phones home

Washington Despite being over 7 billion miles from home, the small NASA spacecraft Pioneer 10 has managed to send a signal back to its ground controllers — ending speculation that it had finally fallen silent.

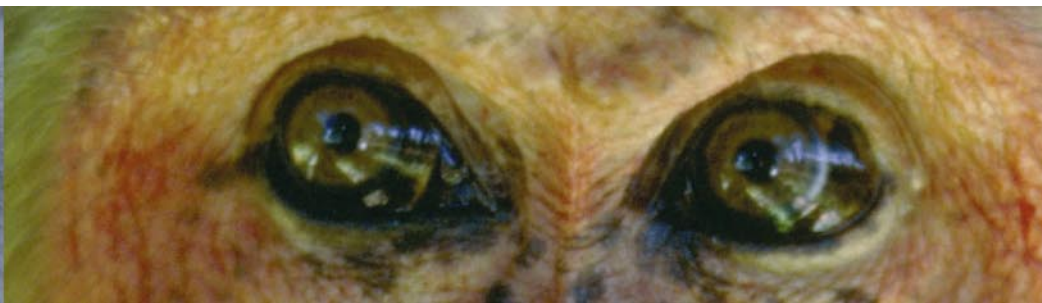
NASA scientists had heard nothing from the 29-year-old spacecraft since last summer. But a signal sent up from a radiotelescope facility in Madrid, Spain — a test of communications technologies for future missions — was received by Pioneer 10, which then returned a signal to Madrid.



Pioneer 10: sending messages from beyond.

Pioneer 10 was the first spacecraft to make direct observations of Jupiter, and finished its mission in 1997. It is currently heading towards the Taurus constellation — but will take 2 million years to reach the nearest of its stars.

AP



Decisions, decisions...

By recording the electrical activity of individual neurons in monkeys, neuroscientists are beginning to understand how the brain makes simple decisions. Bas Kast considers the links between perception and action.

"As soon as questions of will or decision or reason or choice of action arise, human science is at a loss."

When the US linguist and political scientist Noam Chomsky uttered these words in a television interview in 1978, they had the ring of truth. Researchers were starting to get to grips with how the brain perceives events and how it acts upon them. But they had no real understanding of the decision-making processes that link perception and action.

Since then, neuroscientists interested in vision and other sensory systems have studied the brain areas involved in various levels of information processing. Meanwhile, other researchers have unravelled the mechanisms behind complex body movements. "The groups have met at the missing link — the brain locus of decision," says William Newsome, a neuroscientist at Stanford University in California. And over the past few years, Newsome's group, plus a handful of others, has started to reveal the neural mechanisms that underlie simple decision-making.

By stripping decision-making down to its bare essentials, Newsome hopes to reveal the basic mechanisms at work in the brain. "This terrain is sufficiently complex even for the simplest cases," he says.

In a landmark 1996 experiment¹, for instance, Newsome's group presented monkeys with a set of moving dots, a fraction of which were moving in the same direction. The monkeys were trained, using rewards of a drink, to work out which way most of the dots were moving and make a quick, jerk-like, eye movement in the same direction. The task was easy when all the dots were moving together. But in some trials only a few dots moved coherently in one direction, while the others moved around randomly.

The decision became more difficult as the proportion of coherently moving dots decreased. And in the extreme condition, where all dots were moving at random, the monkeys were forced to guess.

Spot the difference

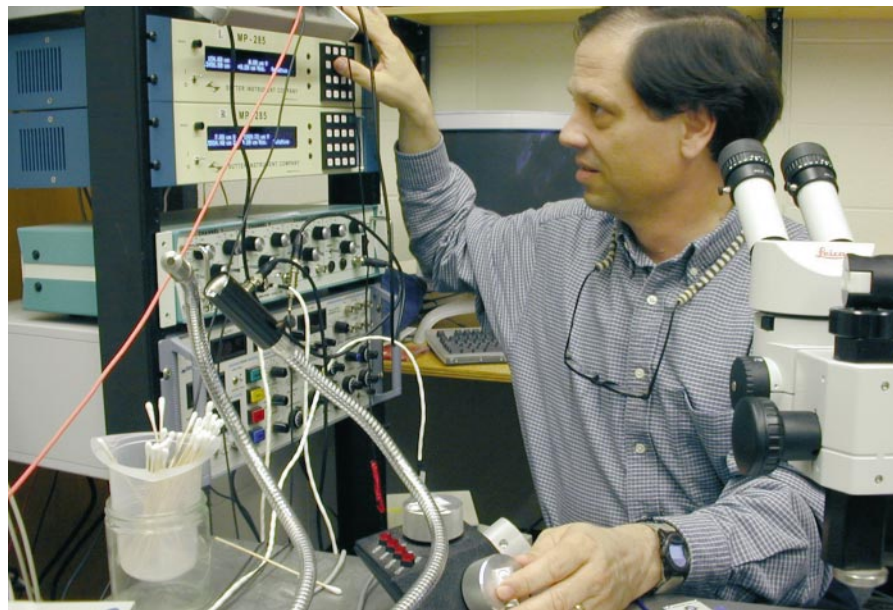
To discover how each monkey used the moving dots to make a decision, Newsome recorded signals — voltage pulses — sent by individual neurons in the monkey's brain while it was working on the task.

Neuroscientists were already familiar with neurons involved in visual processing that are 'tuned' to motion in a particular direction — these cells only fire when they see objects moving in their preferred direction. Similar cells in motor areas fire when the monkey moves its

eyes or limbs in a particular direction.

Together with Michael Shadlen, Newsome recorded the activity of cells in the lateral intraparietal cortex (LIP), a brain region linking areas involved in visual processing to those that control eye movement. They identified LIP cells that responded either to dots moving to the left or to the right. But unlike the tuned cells in visual processing areas, the LIP cells also fired when the monkeys were about to make a guess about the direction in which randomly moving dots were going. Clearly, they were not simply responding to the pattern. "After a while, we noticed that we could predict the direction of the monkey's eye movement by monitoring the activity of a single cell alone," says Newsome.

These cells could merely have been



First choice: William Newsome has identified brain cells that appear to make simple decisions.



CORBIS

controlling the monkeys' eye movements. But Newsome and Shadlen noticed that the cells fired off more pulses when the decision was easy. As the number of coherently moving dots decreased and the choice got harder, tuned cells responded with fewer pulses, even though the monkey was still making a decision to move its eyes. The cells seemed both to respond to the difficulty of the pattern and to influence the monkey's movement, something a purely visual or motor cell could not have done. "Rather than reflecting either sensory or motor information, these cells seemed to integrate both," says Newsome.

Newsome believes that his cells are 'decision neurons', and other groups are now building from his work. After moving to the University of Washington in Seattle, Shadlen teamed up with neuroscientist Jong-Nam Kim to try to explain how such cells make the choice of whether to tell the motor system to make an eye movement.

Under observation

In 1999, Kim and Shadlen identified a similar population of decision cells in a region near the front of the brain called the prefrontal cortex². These neurons are connected to a brain region involved in visual processing called the middle temporal (MT) area, which contains cells that respond to either leftward or rightward movement. Kim and Shadlen suggested that the decision neurons make their choices by monitoring the activity of the direction-tuned MT cells, and then send instructions to cells in the superior colliculus that control movement.

If a monkey looks at a pattern in which most of the dots are moving to the right, neurons in the MT tuned to rightward movement will become active. Kim and Shadlen reasoned that decision neurons observing this burst of activity, together with a corresponding lack of activity in left-tuned MT neurons, will activate the motor system to jerk the eyes to the right so that the monkey receives its reward.

Shadlen and Kim's model has yet to be confirmed by recording from several cells simultaneously. But it does explain why decision neurons fire less as the number of coherently moving dots decreases, as the difference between the firing rates of the left- and right-tuned MT neurons will become less marked.

Decision cells could form the basis for a 'central executive' in the brain.

The model also explains why monkeys sometimes make the wrong decision. Like other neurons, MT cells are inherently 'noisy' — the number of times they fire varies from trial to trial, even when the monkey sees exactly the same stimulus. This means that a burst of firing by left-tuned cells could, by chance, occasionally overwhelm a low signal from right-tuned cells, even if the majority of dots are moving to the right. But Kim and Shadlen's model says that it should still be possible to predict the direction of the monkey's eye movement from the firing of decision cells. This is exactly what Newsome and Shadlen had found previously¹.

The study of decision neurons has now extended beyond vision. Ranulfo Romo, a neurophysiologist at the National Autonomous University of Mexico, has identified neurons involved in choices based on tactile stimuli³. Similar cells for decisions based on smell or sound may also exist.

Executive control

But Romo and some other neuroscientists go one step further. They propose that the brain's frontal lobe contains decision-making cells that are independent of any of the senses. Such cells could form the basis for a 'central executive' in the brain — an area in overall control of decisions. Central executives are a feature of many psychological models of decision-making and the idea gels naturally with our feeling of being in charge of our bodies and minds.

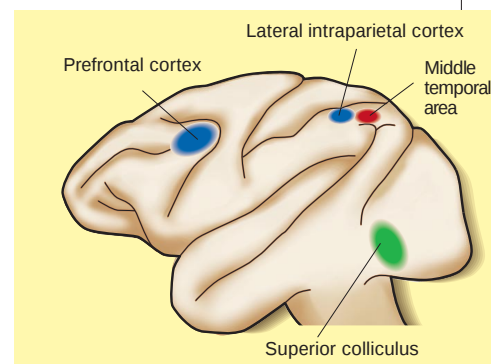
But Newsome and other experts question whether such a central decision-maker will ever be found. Complex decisions, Newsome suggests, are made in a distributed way. "The locus of decision lies in the pattern of connections between sensory inputs and motor outputs. Just as we cannot pinpoint the self in the brain, we cannot determine the decision-maker," he says. "But opinion is highly divided. At the end of the day, we are all just following our intuitions on this issue."

Determining who is right would involve training monkeys to make decisions based on information from more than one of their senses. And so far, such sophisticated experiments have proved beyond the reach of even the most skilled neurophysiologists.

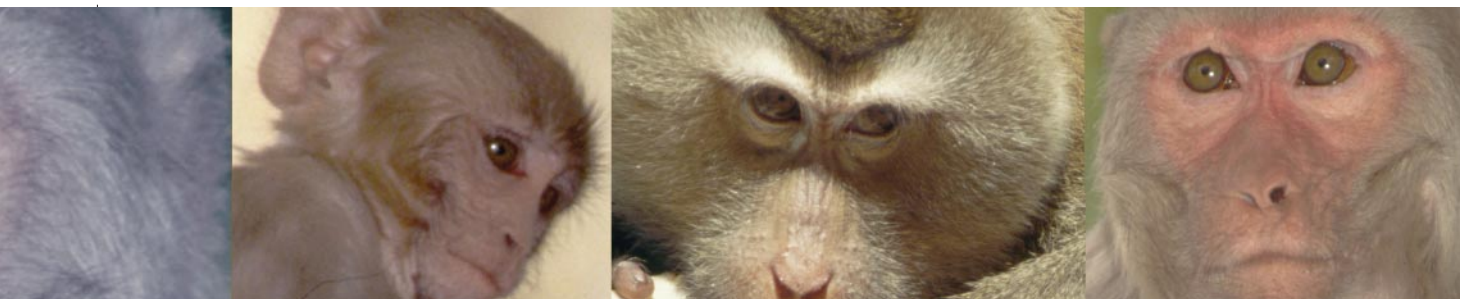
Other neuroscientists believe that the new ideas about decision neurons can be extended to processes that seem not to involve conscious control. Confronted by an overwhelming mass of visual information, areas of the brain involved in visual processing constantly make 'decisions' that simplify what we see. Researchers can investigate how the brain does this by using stimuli that make the decisions difficult. The Necker cube illusion is one example. Our perception of the cube seems to jump between two states — in the version shown overleaf, the green dot can be either on the front or back face of the cube. Unable to decide which is correct, the brain switches back and forth between the two.

Neuroscientists David Leopold and Nikos Logothetis of the Max Planck Institute for Biological Cybernetics in Tübingen, Germany, have used a similar phenomenon to study perceptual decisions. Known as binocular rivalry, the effect can be induced by projecting a different pattern into each of a subject's eyes. Even though both patterns are present at the same time, human volunteers report seeing each one alternately⁴.

Leopold and Logothetis have also measured the activity of neurons in visual areas of brains in monkeys that were viewing different patterns through each eye — one



Cells that control eye movement (green) are given instructions by decision-making cells (blue), apparently based on the latter's observations of cells in the visual-processing region (red).



moving upwards, the other downwards⁴. The monkeys had previously been trained to pull a different lever in response to each pattern, so by recording the animals' responses, the researchers could tell which pattern their brain was perceiving at any one time.

As expected, Leopold and Logothetis identified visual-processing cells tuned exclusively to either upward or downward motion. But the activity of other cells in the monkey's visual processing areas depended on what pattern the monkeys had actually perceived. Some downward-tuned cells, for example, fired only when the monkey pulled on the 'downwards' lever. Leopold and Logothetis propose that these cells are communicating with 'planning' brain areas that are involved in deciding which of the images is actually seen. The mechanism, they argue, is similar to that investigated by Newsome.

Perceptive planning

Other evidence hints that the frontal cortex may be the planning area. The perception of similar illusions can be impaired by damage to the prefrontal cortex⁵, and imaging studies of the human brain have shown that the frontal lobe is active during the perceptual shifts experienced in binocular rivalry⁶. "It doesn't mean that the frontal lobe makes this decision," stresses Leopold, "but deciding to see one image or another, if deciding indeed is the right word to use, might be similar to the decision of moving an eye to the left or the right. The obvious difference is that the output of this behaviour is not a

motor response, but a perceptual change."

Most people intuitively feel that 'deciding' what to see and making conscious decisions are different processes. But Leopold and Logothetis dispute this, pointing out that we can exercise some conscious control over perceptual decisions — try holding one version of the Necker cube steady, for example. And many of the decisions that we make voluntarily, such as braking while driving, become such second nature that they are essentially unconscious actions.

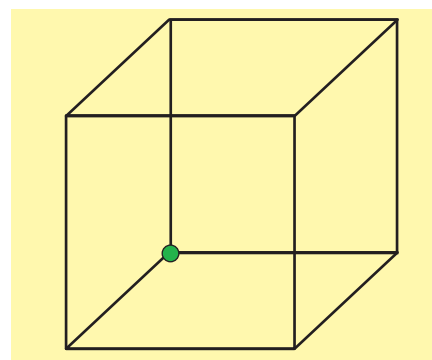
Christian Keyzers, a neuroscientist at the University of Parma in Italy, agrees. "Perceptual decisions are probably the result of a simple, winner-takes-all competition between the neural activity representing alternative perceptions," he says. "It's likely that an equivalent process of neural competition also applies when decisions are not about perceptions but between motor outputs."

But even if the decisions studied by Leopold and Logothetis are equivalent to those investigated by Newsome, both are very different from the endless decisions that fill our waking hours. For one thing, the monkeys studied by neuroscientists are responding to simple stimuli, whereas most of our conscious decisions take account of accumulated knowledge or experience.

Remembrance of things past

Michael Platt of Duke University in Durham, North Carolina, and Paul Glimcher of New York University have attempted to address this objection. They trained monkeys to focus on a light and, when the light changed colour, to make an eye movement. In one block of 100 trials, an eye movement to the right would be rewarded with more than twice as much orange juice as one to left. In other blocks, the rewards were reversed. Over the course of each block of 100 trials, the monkeys learnt which direction brought the biggest reward and consistently moved their eyes in that direction. Lacking a stimulus to tell them which way to go, the monkeys based their decisions on the reward associated with previous trials.

Platt and Glimcher recorded data from LIP neurons tuned to eye movements and found that right-tuned cells fired at higher frequencies when the monkey could expect a greater reward by going to the right⁷. This knowledge had to come from the monkey's



Boxed in: the Necker cube illusion confuses our brain as it cannot work out whether the green dot is at the front or the rear of the object.

memory of previous trials, suggesting that the LIP cells were using this memory to help the monkey make its decision. "This is the kind of problem an animal has to solve all the time," says Platt. "Most of our choices are guided by the pay-offs associated with different actions in the past." He is now searching for the brain region that signals the size of the reward to the LIP neurons.

For the different groups studying the neurophysiology of decision-making, expanding their work to incorporate more complex decisions is the next big step. Even the most sophisticated of the decisions studied in current experiments are still a long way from simple everyday decisions such as what to eat for dinner. The big decisions in our lives — such as whether to take a new job or to get married — are still as inaccessible as Chomsky said they were in 1978.

Newsome does not deny this. But he believes that starting simple provides the best means to investigate the processes that underlie more complex decisions. "Maybe the kinds of decisions we presently study are not what we typically regard as the most interesting decisions that we make every day," he concedes. "But, hey, the neuroscientist has to start somewhere."

Bas Kast writes for *Der Tagesspiegel* in Berlin.

1. Shadlen, M. N. & Newsome, W. T. *Proc. Natl Acad. Sci. USA* **93**, 628–633 (1996).
2. Kim, J.-N. & Shadlen, M. N. *Nature Neurosci.* **2**, 176–185 (1999).
3. Romo, R. & Salinas, E. *Annu. Rev. Neurosci.* **24**, 107–137 (2001).
4. Leopold, D. A. & Logothetis, N. K. *Trends Cogn. Sci.* **3**, 254–264 (1999).
5. Ricci, C. & Blundo, C. *Neuropsychologia* **28**, 1163–1173 (1990).
6. Lumer, E. D., Friston, K. J. & Rees, G. *Science* **280**, 1930–1934 (1998).
7. Platt, M. L. & Glimcher, P. W. *Nature* **400**, 233–238 (1999).



Look and learn: Michael Shadlen is studying how decision cells make their choices.

Private investigations

The tortuous tale of a probe into charges of scientific misconduct levelled at a rising neuroscientist raises questions about the adequacy of US procedures to tackle the problem. Rex Dalton reports.

Four years ago, when Harvard University officials started to investigate allegations of scientific misconduct against neuroscientist Evan Dreyer, he quickly admitted fabricating some of his data. But despite this confession, the case only came to a close last November, when Dreyer agreed to an unusually heavy punishment: a 10-year debarment from receiving federal research funds.

During the investigations, Dreyer moved from Harvard to the University of Pennsylvania, and won a new federal grant. What's more, a clinical trial of a drug was initiated that relied, in part, on a further study conducted by Dreyer that was thrown into question as a result of the probe — for which government officials have never found the primary data.

Federal officials responsible for policing scientific misconduct are concerned that the case exposes serious flaws in current US procedures for investigating misconduct in biomedical research. "I think this whole event with Dreyer has opened a lot of eyes on a lot of issues," says Ron Geller, acting research integrity officer at the National Institutes of Health (NIH) in Bethesda, Maryland.

The making of a monster

Scientific misconduct has been a highly contentious issue in the United States ever since the late 1980s, when a congressional committee started to investigate allegations against scientists who had worked with such prominent figures as Nobel laureate David Baltimore, now president of the California Institute of Technology. In March 1989, the NIH created the Office of Scientific Integrity (OSI), which was given wide-ranging investigative powers.

But in the years that followed, many of the OSI's investigations failed to demonstrate misconduct — including the case in which Baltimore had become embroiled. Eventually, the federal government responded to complaints from scientists that the OSI had become an Orwellian monster, and handed primary responsibility for investigating misconduct back to universities and medical centres. The revamped Office of Research Integrity (ORI), formed in May 1992 within the NIH's parent body, the Department of Health and Human Services, was limited to a supervisory role, reviewing the outcome of institutional probes and administering sanctions.

But an examination of the Dreyer case

The effectiveness of the current system ought to be the focus of concern.

suggests that this decentralized system can leave many issues unresolved. If a researcher under investigation switches universities, for example, the new institution is likely to be kept in the dark. Investigations are also liable to drag on for years as the defendant uses a wide array of available defensive tactics. In the interim, questionable data can remain unchallenged, even as the investigation throws doubt on their veracity. And such data can continue to be used to support clinical trials on patients.

The Dreyer case ended only after he had appealed against an ORI finding of misconduct. Although he finally agreed to the debarment, and conceded that the ORI could demonstrate misconduct, he refused to admit to all of the ORI's allegations regarding data fabrication, or to cooperate with the ORI's reviews of the data. Dreyer also declined to be interviewed for this article, issuing a statement claiming that "Harvard failed to conduct a good-faith investigation".

The investigation began in early 1997, when Dreyer was a Harvard associate professor at the Massachusetts Eye and Ear Infirmary (MEEI). Dreyer was studying the neurotoxic effects of the excitatory amino acid glutamate in the eye. But it was in con-

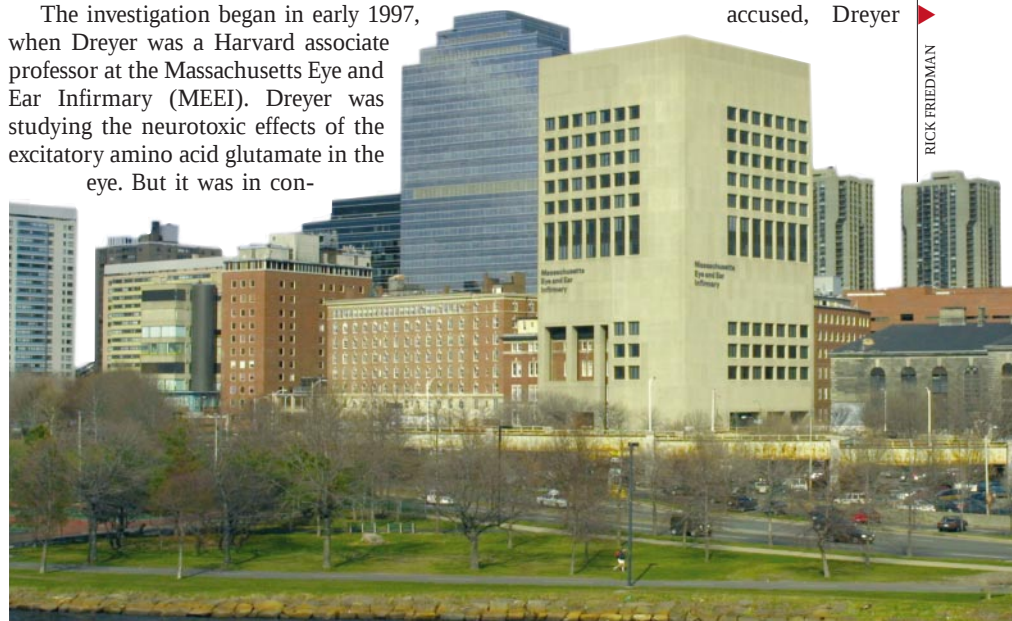
nection with his studies of glutamate's effects in the ear that allegations of misconduct first arose. The study used guinea pigs as an animal model of Meniere's disease, a disorder of the inner ear in which a build-up of fluid causes dizziness, hearing loss and ringing in the ear.

Data discrepancies

When this study was first questioned by other MEEI faculty, Dreyer submitted a computer disk purporting to show high-performance liquid chromatography (HPLC) data for glutamate levels in guinea pigs. But analysis by other scientists at the MEEI determined that the data on the disk were probably fabricated. Shortly afterwards, Dreyer admitted to Harvard officials that he had faked the HPLC results.

University officials continued with a full investigation, following the trail of data in an unpublished manuscript, in an abstract for a meeting, and in applications for an NIH grant and an award for mentoring junior scientists. Dreyer then began an aggressive defence. Deviating from his original confession, he accused a co-worker of faking the HPLC data, and claimed that the investigation was in retaliation for his attempts to expose alleged fraud in billing for surgery by other MEEI faculty. Harvard officials found that both allegations were without merit.

Then, three months after being accused, Dreyer



Under scrutiny: the misconduct probe rocked Harvard's Massachusetts Eye and Ear Infirmary.

RICK FRIEDMAN

announced that he was leaving Harvard to assume an appointment as co-director of the glaucoma service at the University of Pennsylvania's renowned Scheie Eye Institute in Philadelphia. The confidentiality of Harvard's probe into Dreyer meant that Pennsylvania officials were not informed of the misconduct allegation when they hired him.

Highly recommended

Robert Barchi, provost of the University of Pennsylvania, says that its hiring procedures did not include questions that would have required Dreyer to reveal whether he was the subject of a misconduct investigation. Dreyer was given "extraordinarily positive" recommendations by eight neuroscientists nationally, Barchi told *Nature*, including two from Harvard.

In August 1998, just over a year after Dreyer had moved to Pennsylvania, Harvard completed its investigation and concluded that he had fabricated data. But Dreyer was uncooperative towards the subsequent ORI inquiry, which as a result took until April 2000 to complete, reaching the same conclusion. In the interim, Dreyer sued the MEEI over the alleged surgical billing fraud, only dropping the case at the final settlement of his appeal against the ORI ruling last November.

Two weeks before the April ORI ruling, Dreyer obtained a new NIH grant — a five-year award totalling more than \$1.25 million — for research on the role of excitatory amino acids in glaucoma, in which a build up of fluid in the eye damages the optic nerve. According to the confidentiality rules under which the ORI operates, the NIH officials handling the grant application were not aware of the misconduct investigation. So when the ORI delivered its verdict, they were faced with awarding a grant to someone who had already been found guilty of misconduct by a leading research university. The officials then worked out a deal with the Pennsylvania authorities to shift the funds to another

researcher, Alan Laties, on the understanding that he would check the data underlying the grant application before proceeding with the work.

This checking was deemed necessary because the Harvard investigation had also recommended an internal administrative review of other research data, which Dreyer and co-authors had used to publish an article on glutamate exacerbating glaucoma in humans and monkeys (*Arch. Ophthalmol.* **114**, 299–305; 1996). Primary HPLC data for this article have never been found, say federal officials familiar with the case.

In January 2000, the ORI became sufficiently concerned about this paper that it notified both the NIH and the Food and Drug Administration (FDA). The FDA was informed because ORI officials believed that the paper had been used in support of a clinical trial of a drug called memantine being undertaken by Allergan of Irvine, California. Memantine appears to be protective against glutamate's neurotoxic effects, and has been used in Germany for some time to treat dementia.

With Dreyer gone from Harvard and being uncooperative, the university's administrative review never got off the ground. Laties says that he was not informed of Harvard's concerns regarding missing data for the glaucoma paper, either directly or through his superiors at the University of Pennsylvania. And although ORI and FDA officials did meet to discuss the matter, it remains unclear what happened as a result — commercial confidentiality means that FDA officials cannot discuss the issue.

In an interview with *Nature* in December, Larry Wheeler, Allergan's chief scientific officer, said that no one from the FDA had

LABORATORY SCIENCES

Elevated Glutamate Levels in the Vitreous Body of Humans and Monkeys With Glaucoma

Evan B. Dreyer, MD, PhD; David Zurakowski, PhD; Robert A. Schumer, MD, PhD; Steven M. Podos, MD; Stuart A. Lipton, MD, PhD

Objective: To explore the possibility that the excitatory amino acid glutamate might be associated with the disease process of glaucoma, which is characterized by the death of retinal ganglion cell neurons and subsequent visual dysfunction.

Methods: Amino acid analyses were performed on vitreous specimens that were obtained from patients who were undergoing cataract extraction. Samples were collected prospectively from those patients who sustained inadvertent rupture of the posterior capsule between 1988 and 1993. An additional set of specimens, obtained from both eyes of monkeys, was analyzed; in these monkeys, glaucoma had been experimentally induced in one eye only.

Results: A twofold elevation in the level of glutamate was detected in the vitreous body of the group of patients with glaucoma when compared with that in a con-

trol population of patients with cataracts only. An even greater elevation of the glutamate level was found in the vitreous body of glaucomatous eyes of monkeys when compared with that in control eyes. No statistical differences were detected among other amino acid levels from the vitreous body of glaucomatous and nonglaucomatous eyes in humans or monkeys.

Conclusion: The vitreous body of glaucomatous eyes contains elevated levels of glutamate. This finding suggests that glutamate may be involved in the pathogenesis of glaucoma. (Arch. Ophthalmol. 1996;114:299–305)

Objective: To explore the possibility that the excitatory amino acid glutamate might be associated with the disease process of glaucoma, which is characterized by the death of retinal ganglion cell neurons and subsequent visual dysfunction.

Methods: Amino acid analyses were performed on vitreous specimens that were obtained from patients who were undergoing cataract extraction. Samples were collected prospectively from those patients who sustained inadvertent rupture of the posterior capsule between 1988 and 1993. An additional set of specimens, obtained from both eyes of monkeys, was analyzed; in these monkeys, glaucoma had been experimentally induced in one eye only.

Results: A twofold elevation in the level of glutamate was detected in the vitreous body of the group of patients with glaucoma when compared with that in a con-

Missing: data for this paper have not been found.

ever contacted him about the missing data. He referred inquiries to neuroscientist Stuart Lipton, senior author on the glaucoma paper. Lipton, formerly of Harvard but now at the Burnham Institute in La Jolla, California, is named with Dreyer as inventor on a patent for using memantine to treat glaucoma. Lipton says the questioned data are Dreyer's responsibility. Lipton told Harvard officials in October 1998 that he did not have the primary data. But unless someone makes a specific allegation of scientific misconduct regarding data in the glutamate/glaucoma paper, Harvard's probe will go no further.

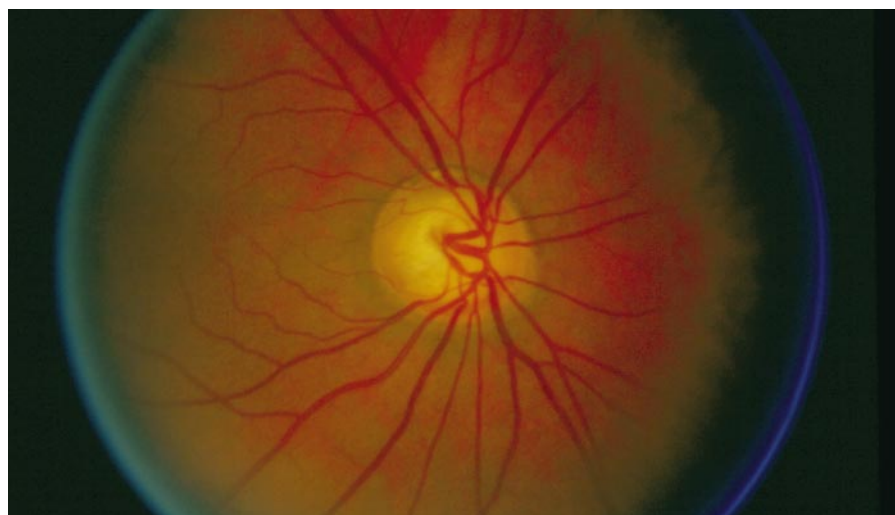
Wheeler says that the ongoing clinical trial is underpinned by several animal studies. But the paper co-authored by Dreyer and Lipton is regarded as the seminal publication linking glutamate with glaucoma in humans.

Wiley Chambers, an official at the FDA's eye-drug evaluation centre, says that the glutamate/glaucoma paper will be subjected to closer scrutiny if Allergan seeks approval to use memantine as a treatment for glaucoma after its trials are complete. Again, because of commercial confidentiality, the FDA will not discuss its decision to let the trial go ahead.

Scientists and some university administrators are now expressing concerns about the costs involved in conducting misconduct investigations, which they would like the government to share. When extensive reviews of data are required and the accused scientists are uncooperative, the costs can spiral.

To the federal officials who monitor misconduct cases, these complaints are frustrating — particularly as they are coming from many of the same people who previously lobbied against the centralized OSI system. Pointing to the loose ends left hanging by the Dreyer case, they argue that it is the effectiveness of the current system, rather than its cost, that ought to be the focus of concern. ■

Rex Dalton is *Nature's* West Coast US correspondent.



One in the eye: glaucoma can be diagnosed by observing the retina.

Science should help teach children the meaning of humanity

Sir — The scientific community must push forward new curriculum ideas before more US school boards like those in Kansas and Kentucky attempt to deny their students a proper science education¹. Rapid changes in the world economy and in social structures are prompting the need for individuals who can adjust and thrive in a changing world. These challenges require educational innovation. Yet according to international test scores, US public schools are stagnating.

In 1963, the view emerged that “the rate of change in the society in which we live forces us to redefine how we shall educate a new generation”². Environmental problems, population growth, violence and growing suburbanization were placing increasing demands on resources. These demands prompted the creation of a curriculum designed to use education, rather than government regulation and penalties, to improve the human condition².

The result was Macos (Man: a course of study), a school programme integrating anthropology, biology and the latest concepts of educational psychology. Developed by teachers and specialists in the areas studied, with the help of a grant from the National Science Foundation (NSF), it stayed in use for more than a decade. At its peak the Macos curriculum was taught in 47 states and 1,700 schools, and was hailed as significant educational progress³. Forty years later, the same social and economic problems remain. Perhaps it is time for the social and biological sciences again to be made central to a multidisciplinary curriculum based on complex systems of human societies and behaviour.

Macos attempted to achieve social change by the use of anthropological and biological models designed “to help children understand what it means to be human”² by addressing basic questions about humanity. Five goals were established: to give pupils respect for and confidence in the powers of their own minds; to use this to give them power to think about the human condition and society; to provide workable models to analyse the nature of society and the human condition; to impart respect for the capacities and humanity of the human race as a species; and to provide a sense of the unfinished business of human evolution².

Macos was a daily 40–45-minute lesson for fifth- and sixth-grade students (aged 9 to 11). The course included studying the life and culture of the Netsilik Eskimos and

their habitat. By examining the biological and anthropological systems of the Eskimos, children could examine their own culture and habitat⁴. The course was discontinued in 1976 by the NSF, prompted by Senator John Conlan (Republican, Arizona), who objected to it because it dealt with topics such as evolution, reproduction and violence⁵.

In essence, Macos was a beginning course in philosophy, using several disciplines to address basic questions of humanity, focusing on creating complex questions rather than answering them. Fact acquisition was secondary to establishing a foundation on which learning and thinking would grow. While the idea of creating questions instead of delivering facts may be frowned on by proponents of standardized tests, it is essential for teaching children to think.

An educational curriculum needs to be designed today that has the same character and purpose as Macos, but based on broader, complex systems, with computer databases and the Internet used as integral tools by teachers. The development of the curriculum must be a collaborative effort by leaders of several disciplines to provide a comprehensive and successful educational programme⁶.

Scientists must not sit back and watch movements towards irrationality, such as the judgements in Kansas and Kentucky in 1999. A united scientific community can provide an impetus for future social change by introducing and supporting a strong science curriculum based on complex systems.

Greg Kontos

Information Technology Department, College of William and Mary, Jones Hall, Room 105, Williamsburg, Virginia 23185, USA

1. *Nature* **402**, 847–848 (2000).
2. Bruner, J. S. *Toward a Theory of Instruction* 23–101 (Belknap, Cambridge, MA, 1966).
3. Nelkin, D. *The Creation Controversy: Science or Scripture in the Schools* 51 (Norton, New York, 1982).
4. Weber, G. *Phi Delta Kappan* **57**, 79 (1975).
5. Dow, P. *Phi Delta Kappan* **57**, 81 (1975).
6. Clark, J. W. in *Introduction to Systems Philosophy: Toward a New Paradigm of Contemporary Thought* (ed. Laszlo, E.) 301 (Harper Torchbooks, New York, 1972).

Particle physicists need a common objective

Sir — I would like to comment on two points arising from your Opinion article and News story on the the TESLA project (*Nature* **410**; 395 & 397; 2001).

First, the cost of 3,136 million euros (US\$2,811 million) for the TESLA 500-GeV electron–positron linear collider is based on experience of construction of a complete test facility, on operation for four years, and on a full costing performed by

industry, including the effects of mass production.

I disagree with your opinion that the cost will double when labour and operation costs are included, because all components will be fabricated by industry and the associated labour costs are included in the price. Testing and assembly of components will require roughly 7,000 person-years, the cost of which will have to be added to the investment cost, and which amounts to no more than 10–20% of the total cost, depending on national salary scales. Operating costs are not included in any of the existing cost estimates: for TESLA these should be 120 million euros a year, or 3% of the investment.

Second, you state in your editorial on the German, US and Japanese plans for linear accelerators that Europeans should accept that they have had their turn with CERN's Large Hadron Collider. Your News story says that TESLA in Germany “could mark the virtual eclipse of particle physics in the United States and Japan for a generation”. In my view, regional balance is good but not vital for a field. This has been demonstrated successfully by astronomers who go where the sky tells them, without endangering national programmes.

In particle physics we need a new model of international collaboration to be able to build and operate accelerators as global endeavours, in which all partners have shares and a say, thus making a facility for all countries, independent of the location. It is important that particle physicists everywhere get behind one common project, and convince their governments to fund and build it wherever the strongest political support is found.

Albrecht Wagner

DESY, D-22603 Hamburg, Germany

No campaign to strip Baltimore of his Nobel

Sir — In his review of *Ahead of the Curve*, Shane Crotty's biography of David Baltimore (*Nature* **410**, 746; 2001), Robert Bazell perpetuates a misunderstanding of my role in the ‘Baltimore case’ when he writes that I “campaign to have Baltimore's Nobel prize rescinded and to have him expelled from the National Academy of Sciences”. The assertion that I led a campaign to destroy Baltimore is false. Its origin may lie in the reappearance at that time of the long-discredited rumour that David knew of Howard Temin's discovery of avian reverse transcriptase before he isolated its murine equivalent — I had inappropriately repeated what turned out to be an unfounded story.

Earlier, before the Dingell congressional hearings under which David was investigated for his support of a colleague accused of scientific fraud, I had in fact tried to defuse the situation, first by holding a small meeting on scientific integrity at our Banbury Center to try to review what was going on. Although invited, David declined to come, but his close associates did. Second, I went to the offices of Congressman Norman Lent, Long Island's Republican counterpart to John Dingell, to tell him that the controversy involved not fraud but possible mistakes of scientific judgment, and as such it should not merit the attention of a congressional committee.

I have only the highest respect for Baltimore as a scientist and leader of science, but his intransigence at the time made many of us worry that the affair was harming not only himself but US science as well. I in no way feel apologetic for not supporting behaviour that I feel is antithetical to the scientific tradition in which I was raised. The Harvard professors who saw the need to publicly question David's behaviour only reluctantly joined the fray. In reporting their and my behaviour as "horrid", Bazell does us a great injustice. Under no circumstances could John Edsall or Paul Doty, say, be accused of jealousy of a peer's meteoric career. I and the Harvard professors were deeply bothered by what seemed to us actions not appropriate for an individual of such talents. I totally concurred in my former colleagues' forthright wish for justice to prevail.

James D. Watson

*Cold Spring Harbor Laboratory, PO Box 100,
Cold Spring Harbor, New York 11724, USA*

How inbreeding affects productivity in Europe

Sir — Navarro and Rivero in Correspondence¹ for the first time quantified favouritism ("inbreeding") in Spanish universities, showing that it is at least 10 times higher than in France, the United States and the United Kingdom. I have now quantified this phenomenon in 51 universities from 14 European countries.

I examined the relationship between the percentage of papers published by each country and its level of inbreeding. I gathered information from 51 ecology or zoology departments via a brief questionnaire requesting information on the number and rank of teaching positions, as well as the number of these staff trained at the same university.

I obtained responses from at least two universities in each country. I estimated "inbreeding" as the percentage of staff in these positions trained at the university. I found significantly more variance in inbreeding among than within countries ($F = 12.38$, d.f. = 13, 37, $P < 0.0000001$; one-way analysis of variance), showing that countries are consistent in their degree of favouritism.

Spain registered the second highest level of inbreeding, averaging 88%, surpassed only by Portugal (91%). Figures for the remainder are (in %): Italy (78), Austria (73), France (65), Norway (56), Belgium (52), Finland (48), Netherlands (40), Denmark (39), Sweden (32), Switzerland (23), United Kingdom (5.2) and Germany (1).

To explore the relationship between level of inbreeding and scientific productivity, I used May's enlightening report², in which he assesses the quality of the contribution of various countries to world scientific knowledge. May provides information about the share of papers and citations in science, medicine and engineering provided by 15 countries, not including Portugal, Spain, Austria, Norway and Belgium. If, conservatively, I assign these countries a value of 0.5, a significantly negative correlation appears between the percentage of papers published by each country and its level of inbreeding ($R_s = -0.60$, $P = 0.02$, $n = 14$). That is, overall scientific productivity correlates negatively with the percentage of inbreeding.

In the 1980s the Spanish government attempted to end inbreeding in its universities by the University Reform Law (LRU). The Real Decreto 1888/1984 of 26 September, which regulates the employment system in Spanish universities, clearly states: "The research activities of the candidates shall be evaluated as the priority merit" (Art. 8.2.a). This reform has, however, failed to change the system³; indeed, the problem has become even more serious, as pointed out in two *Nature* editorials^{4,5}.

The inbreeding system is extremely stable. The solution needs vigorous measures: first, every position should be advertised internationally; second, there should be no local members on appointment committees; and third, lecturers or full professors with low scientific productivity should not serve on committees that appoint professorships. As things stand, Spanish universities are autonomous and they do not want the system to change because, in general, university politicians (people who manage universities) are not good researchers and do not consider

that scientific productivity is paramount.

When will the Spanish government decide to implement the necessary drastic reforms?

Manuel Soler

*Department of Animal and Ecological Biology,
Faculty of Sciences, University of Granada, 18001,
Granada, Spain*

1. Navarro, A. & Rivero, A. *Nature* **410**, 14 (2001).
2. May, R. M. *Science* **275**, 793–796 (1997).
3. Rodilla, V. *Nature* **376**, 290 (1997).
4. *Nature* **389**, 767 (1997).
5. *Nature* **396**, 709 (1998).

Philanthropists are paying their dues

Sir — I would like to add some points to your excellent News Feature on biomedical philanthropy (*Nature* **410**, 140–143; 2001), which highlighted the fact that high-risk research not funded by the National Institutes of Health (NIH) is increasingly being supported by philanthropists.

First, the NIH system is one of the best biomedical research funding systems in the world. NIH support of research and research training has catalysed much of the explosion of biomedical knowledge over the past few decades.

Second, the NIH and other US agencies have freely allowed investigators to patent technological discoveries made while they were supported by government funds, a policy responsible for the great success of many entrepreneurs. It is thus only fair that some of their good fortune is funnelled back to support basic research.

Third, over the past few years the NIH has been seeking and supporting high-risk research, for example the National Eye Institute's new R03 programme, the National Institute on Drug Abuse's CEBRA programme and the R21 programme of the National Institute of Mental Health and National Institute of Neurological Disorders and Stroke.

Programme officers from each of these institutes have personally contacted me to inform me of these initiatives and to encourage me to apply. My own lab's high-risk work has received funding from both NIH and philanthropic sources.

Finally, and perhaps most important, the NIH system is only as good as the quality of peer review. If biomedical researchers are unhappy with the quality of the review process, we need only look to ourselves to improve it.

Ben A. Barres

*Department of Neurobiology, Stanford University
School of Medicine, Fairchild D235, 299 Campus
Drive, Stanford, California 94305-5125, USA*

The hidden meanings of maps

Maps can convey far more about the world than the route from A to B.

The New Nature of Maps: Essays in the History of Cartography
by J. B. Harley. Edited by Paul Laxton
Johns Hopkins University Press: 2001. 352 pp.
\$45, £30.50

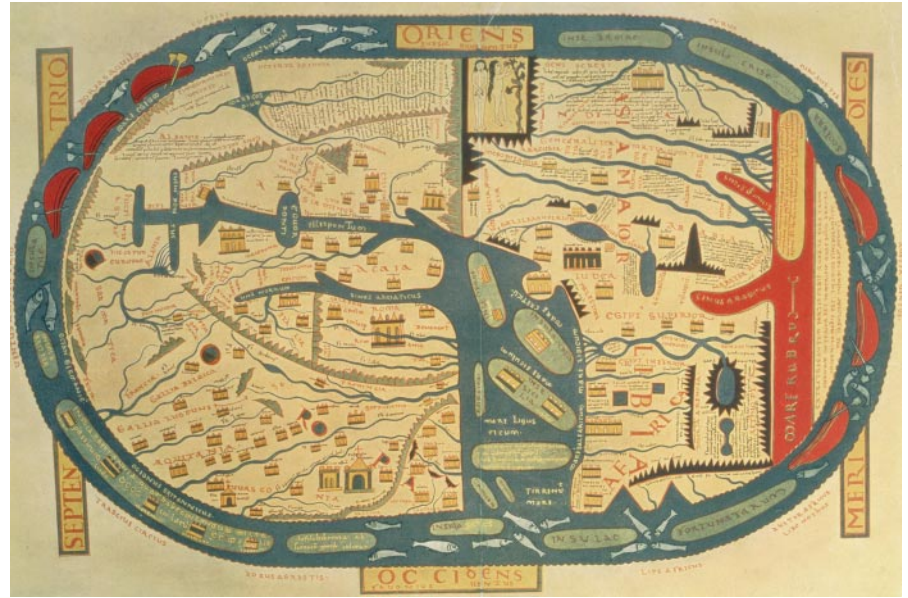
Apollo's Eye: A Cartographic Genealogy of the Earth in the Western Imagination
by Denis Cosgrove
Johns Hopkins University Press: 2001. 352 pp.
\$46.50, £31.50

Catherine Delano-Smith

Few subjects in the humanities have undergone a more rapid and radical turnabout in the way they are studied than has the study of maps, especially pre-modern maps. Non-modern maps were previously characterized in terms of their origin and date, their position in the development of techniques for surveying and production, and their geographical accuracy. Now they are assessed as historical documents. This means that all aspects of a map are critically analysed so as to place it in its proper context, which is as likely to be a context of intellectual history — or the history of literacy or diplomacy — as the traditional context of discovery, exploration and advancement in surveying techniques. No longer, then, need a map be regarded exclusively as a product of science, incorporating astronomical observation and precise measurement, projection and scale.

Although the revised approach goes back to at least 1980 with the publication of M. J. Blakemore and J. B. Harley's *Concepts in the History of Cartography*, the marker for the new history of cartography is commonly set at the first volume of Harley and David Woodward's ongoing *History of Cartography* (University of Chicago Press, 1987). This provided a liberating redefinition of maps as graphic expressions of all manner of spatial distributions produced by all sorts of people, in all sorts of ways and for all sorts of reasons. In *The New Nature of Maps*, written just before his sudden and early death in 1991, Harley enlarges on such topics as 'Maps, knowledge and power', 'Power and legitimation in the English geographical atlases of the eighteenth century', 'Deconstructing the map', 'New England cartography and the native Americans' and 'Can there be a cartographic ethics?'.

Harley's book champions the non-positivist cause. Through his insistence that a society's mapping practice is intimately linked to its culture, he sought to show that maps can have meaning as well as form, and sometimes multiple levels of meaning. They may even have meaning instead of their usual



World views: from the ninth-century evocation by the Spanish monk Beatus ...

narrow practical function. They can convey notions of imperial power, such as in maps created in nineteenth-century Britain, or the silence of the dispossessed — those in colonial America — or of the unrepresented — residents of unmapped urban alleys or rural farmsteads. They can reveal the subtle might of class distinctions, traceable through subscribers to national atlases, or uncover hidden agendas, latent in the way knowledge is presented or not presented. Harley imported his ideas on map history from elsewhere, acknowledging his debt to writers of literary criticism (Jacques Derrida, Michel Foucault, Roland Barthes), art history (Erwin Panofsky, Michael Baxendall) and the philosophy of science (Joseph Rouse).

Harley had chosen the articles to be included in the book and arranged them in the order in which they are presented, but died before he had even made notes for the promised introduction. The task of creating a coherent book out of the amorphous pile of papers on the floor of Harley's study in Milwaukee (where he had moved to from England in 1987) was eventually handed by Johns Hopkins University Press to Harley's literary executor and former research student, Paul Laxton.

With supreme tact, sympathetic insight into Harley's personality and his own deft scholarship, Laxton has produced not Harley's book but a book worthy of Harley. While Harley's enviably fluent and expressive prose stands, Laxton's editing and updating of Harley's innumerable references and cross-references is meticulous and

painstaking. He has also provided a valuable index and an invaluable consolidated bibliography (noting in his Preface that only 53 of the nearly 500 works cited by Harley appear in more than one article and only ten in three or more). Matthew Edney has contributed an updated list of Harley's substantive publications and John Andrews a ruthlessly analytical Introduction.

Andrews' essay is a bonus. One might not agree with all he says, any more than one can always take Harley at face value. But Andrews' cool detachment and informed comment are a counterpoint to Harley's heady espousal of the deconstructionist menu. Harley wrote in the heat of the moment, out of passion for his subject and with a sometimes unguardedly



... to the seventeenth-century depiction by cosmographer Andreas Cellarius.

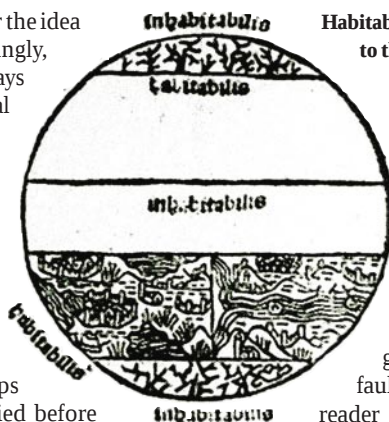
sweeping enthusiasm for the idea of the day. Not surprisingly, his essays don't always stand up to philosophical scrutiny.

Measured consistency is not the point, though; or rather, it was not Harley's objective. He wanted to stimulate and goad readers of early maps, mercilessly and mischievously, into thinking about the maps in different ways. He died before he could leave a blueprint for future work. He wanted, he had told friends, to return to empirical study and book writing, for it was in the maps themselves that inspiration, as well as pleasure, were to be found. But it would have been with his *New Nature of Maps* that he would have urged his readers to take up the intellectual challenge to see early maps with new eyes. Coming from Harley, the challenge is irresistible.

Denis Cosgrove's book presents a different kind of journey, less theoretical and much more specific. Cosgrove comes from the academic generation after Harley's and can take the new history of cartography as read. Maps and globes, and their representation in art and cartography, have symbolic meaning for Cosgrove. His concern in this book is to show how images of the globe "have constructed and communicated the distinctive Western mentality".

He starts with the Classical world view (Apollo the Sun god, whose gaze over the Earth is "synoptic and omniscient [and] intellectually detached") and closes with the virtual globe in the age of geopolitics (Apollo the space project). On the way, he touches on the Christian Middle Ages, and deals with the early modern and modern centuries, the oceanic globe depicting the age of discovery, the first printed globes and atlases and (his forte) the globe as emblem in art and cartography. In his later chapters, he sweeps a whole range of topics into his main theme — orreries, longitude, imperial science, universal history, polar exploration and the mapping of continental interiors are just some of the goodies in Cosgrove's cornucopia. There are also echoes of Harley's themes: power in various guises and propaganda, for example.

Cosgrove ranges widely, but with greater firmness of touch on later periods than on earlier ones. His discussions in the medieval chapter seem less well grounded than some longer expositions — for example, the use of globes in the seventeenth century by the Venetian philologist Girolamo Ruscelli (who linked the science of globes to "the moral and metaphysical discourse of emblem and allegory"), and by the Catholic church and Jesuit propaganda (through the symbolism of the cordiform — heart-shaped — map). There



Habitable areas of the globe according to thirteenth-century astronomer Sacrobosco.

are times, too, when the connection between Cosgrove's case studies and their significance for the odd (and misleading) subtitle of "cartographic genealogy" seem slightly forced. Cosgrove's prose style cannot be faulted, though, and the general reader will enjoy the span of his themes. The scholar will find Cos-

grove's notes and references less rewarding than Harley's and the absence of a consolidated bibliography a serious omission.

Catherine Delano-Smith is at the Institute of Historical Research, Senate House, University of London, Malet Street, London WC1E 7HU, UK.

Telling it like it really is

Real Science: What it is, and What it Means

by John Ziman

Cambridge University Press: 2000. 399 pp.

£25, \$39.95

David L. Hull

For the past four decades, John Ziman has been publishing on the nature of science. His latest book is a distillation of all his preceding work, and in it he presents several unexpected positions. First, he argues in favour of the traditional view of science — the norms formulated by R. K. Merton in 1942, such as the universality, communalism, disinterestedness and organized scepticism of science — but disagrees with the view that these govern science throughout its existence. Instead, Ziman claims that these norms characterize only "pure academic science" and only for a relatively short period of time.

Pure academic science flourished from roughly 1834 — when the term 'scientist' was coined — until the 1960s, when this view of science came under attack by relativist, constructivist critics. Since then it has been gradually replaced by 'post-academic science'. It is not that science has remained the same while our views of it have changed, but rather that science itself has undergone basic modification. As a result, our understanding of science must also change.

A second unexpected view is that, for Ziman, 'new' does not imply 'bad'. The new way of doing science can be disconcerting because no one really understands how it works. Although most of Ziman's book concerns academic science, he also explores

post-academic science. He does this by explaining each of Merton's norms and then showing how they have been modified or replaced in post-academic science. For example, according to Merton's norm of disinterestedness, scientists should accurately evaluate the contributions of their contemporaries. This does not mean that scientists are in the least dispassionate in their research, but that academic science is organized in such a way that scientists can best foster their own research by accurately evaluating the contributions of their contemporaries. Unlike many other professions, academic science is reasonably good at policing itself. But in post-academic science, government agencies and industry are playing an increasing role in deciding not only what research should be conducted but also how it should be evaluated. Hence, if post-academic science is to be in any sense 'objective', other norms must come into play.

Ziman knows that his choice of Merton's norms to structure his exposition will lose him one class of readers — those critics who debunk the ethos of science, both past and present, as collective false consciousness. Just as pure academic science was withering away, a variety of critics arose to ridicule the mertonian norms. In fact, these various groups of critics have little in common except their antipathy to Merton. Ziman rejects much of the recent 'science studies', but at least he considers these relativist, constructivist views worth rejecting. Philosophy is quite another matter. In his youth Ziman was "beguiled" by what philosophers had to say about science, but eventually came to see that they were not "telling it like it is". In particular, they ignored the twin facts that science is both social and a profession.

Throughout the book, Ziman presents a litany of instances where philosophers failed, underestimated, were misled and misled others. According to Ziman, philosophers more than any other students of science are responsible for the legendary view of science. This legend has it that "science is magically endowed with an infallible method for achieving absolutely perfect truth". Of course, science has never had these characteristics, nor have any philosophers during the past century or so thought that it did, certainly not philosophers of science. But claiming that one is fighting the forces of evil positivism is too seductive a ploy to eschew. What name, then, can Ziman give his own activity? Relativist constructivists have co-opted the phrase 'science studies', and philosophers have a firm hold on 'philosophy of science'. Thus, Ziman is forced to coin the appellation 'metascience' as a generic term for all types of the study of science.

Ziman's own metascientific views can be characterized as evolutionary and naturalistic. Just as polar bears evolved from brown bears, Einstein's theory evolved from lineal

descendants of Newton's theory. And at the metascientific level, post-academic science has evolved from academic science. The point of these comparisons is to give some credence to science and our views about science. Organisms are adapted to the environments in which they live; if they were not, they would not be here. Similarly, our scientific views about the world in which we live must be reasonably accurate or else the artefacts that we build on the basis of these theories would fail. Our bridges would collapse even more frequently than they do. Traditional philosophers think that our scientific views must be justified, but an evolutionary justification is not good enough. Certainly, no appeal to nature can justify our beliefs about nature. A stronger justification is needed.

Ziman is aware that he is promoting certain philosophical views that traditional philosophers reject, but he seems unaware of at least one implication of his evolutionary view of science — that science cannot be a natural kind. Because both species and science evolve, often quite gradually, neither can be strictly defined. According to the traditional view, natural phenomena can be subdivided into classes that are sharply distinguishable from one another. These classes, or 'natural kinds', must be definable in terms of characteristics that are necessary and jointly sufficient for membership.

The trouble is that although Ziman repeatedly claims that science is a natural kind, one thing that natural kinds cannot do is evolve. Each has its own essence. For example, the natural kind 'triangle' is a closed geometric figure with precisely three sides. Rectangles have precisely four sides. On this account, the whole notion of triangularity evolving into rectangularity makes no sense. One of the major conceptual shifts that Darwin's theory forced on us is that species can no longer be treated as natural kinds but must be construed as historical entities, a special sort of particular. If science is to be viewed from an evolutionary perspective, parallel conclusions follow. Either science is a natural kind and cannot evolve, or else it evolves and cannot be construed as a natural kind. Putting this complaint to one side, for anyone wanting a detailed, realistic, well-rounded view of science, Ziman's *Real Science* is your book. ■

David L. Hull is in the Department of Philosophy, Northwestern University, Evanston, Illinois 60208, USA.

Voyager's treasures lost and found

**The Letters of Sir Joseph Banks:
A Selection, 1768–1820**

edited by Neil Chambers

Imperial College Press: 2000. 420 pp. £33, \$48

John Gascoigne

Joseph Banks is best remembered as the botanist who accompanied Cook on the Pacific voyage of the *Endeavour* and who was instrumental in founding the Royal Botanic Gardens at Kew, west London, in its present form. If pressed, one might recall that he was the man who organized the ill-fated voyage of the *Bounty*. But he was also the longest serving president (1778–1820) of the Royal Society to date and was, in effect, the British government's chief scientific adviser in the critical period of the American, French and Industrial Revolutions. The vast epistolary edifice which records these manifold activities was maintained and classified with great care by the always-organized Banks, but was largely demolished after his death — principally by a distant aristocratic heir seeking a quick profit.

The patient reassembling of this unique record of the scientific and intellectual currents of Britain and the wider learned world in the late eighteenth and early nineteenth centuries was started by Harold Carter, who provides the biographical introduction to this book. It is being continued at the Banks Archive Project at the Natural History Museum, London, where the editor is research curator.

This collection provides a taste of the scholarly banquet that will eventually emerge from the Banks Archive Project. It is intended to convey the range and scope of Banks's many activities, as well as the vivid directness of his epistolary style. For Banks turned his letters into instruments of both persuasion and command, enabling him to hold together his vast empire of correspondents, collectors and collaborators. He generally eschewed high literary polish, but the letters are an art form in their ability to convey succinctly the state of a problem and his proposed response. The correspondent is given a clear sense of his chain of reasoning and, although he is almost uniformly polite,

One of the plants recorded by Banks.

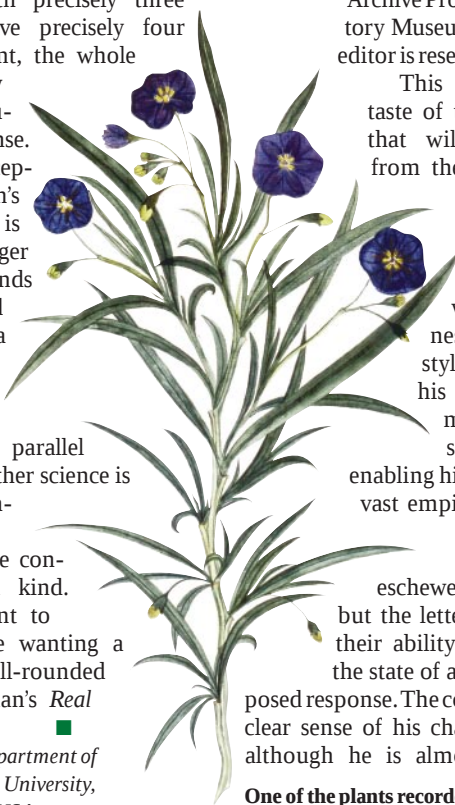


Banks's correspondence is a unique record of the scientific and intellectual currents of his time.

Banks's strong will and assumption of authority are often evident.

This sampling of the more than 20,000 items of Banksian correspondence reproduces 137 of the 6,000 or so letters written by Banks himself. Most aspects of his life are represented: the great Pacific voyage on the *Endeavour* with Cook, the Royal Society and its disputes, his involvement in Lincolnshire affairs as squire of Revesby Abbey, his continuing relations with American and French scientists despite these countries' conflict with Britain, exploration of the Australian and African continents, his relations with the ailing English King George III, scientific advice to government, advancement of the whaling industry, and the organization of the *Bounty* expedition, which was intended to achieve one of Banks's dearest aims — the movement of plants around the world as a form of ecological imperialism, a goal also apparent in his plans for planting Chinese tea in India.

We also see something of Banks in his domestic setting, assuming not only the responsibilities of his immediate family but also those of his wider kin, including a nephew whom he firmly, and characteristically, warned against the "Yawning Gulph of Idleness". The correspondence with Charlotte Seymour, eleventh Duchess of Somerset, with its stoical acceptance of the vicissitudes of old age, comes closest to revealing something of the inner Banks, which remains hidden in the bulk of his letters, devoted to public affairs. Banks's extensive involvement in the affairs of Iceland, which he visited and later helped by alleviating the effects of the British naval blockade during



the Napoleonic Wars, is omitted, but is to be published by the Hakluyt Society.

This selection, then, helps to restore to the map of late eighteenth-century and early nineteenth-century intellectual life a figure too long absent and a great master of the art of letter-writing. It offers a glimpse of the long-submerged epistolary riches that provide such insight into the character of a remarkable individual and his age. ■

John Gascoigne is in the School of History, University of New South Wales, Sydney 2052, Australia.

Mastermind of the bird world

Erwin Stresemann (1889–1972): Life and Work of a Pioneer of Scientific Ornithology
by Jürgen Haffer, Erich Rutschke & Klaus Wunderlich
Deutsche Akademie der Naturforscher Leopoldina: 2000. 465 pp. (German with English summary). DM68, 34.80 euros (to be ordered through G. Thieme Verlag, Stuttgart)
Matthias Glaubrech

Ever since Charles Darwin and Alfred Russel Wallace strove to solve the 'mystery of mysteries', the question of the origin and nature of species has been paramount for evolutionary biologists. Long before the 'evolutionary synthesis' became the theoretical umbrella for systematics, genetics and palaeontology in the 1940s, two of its German architects, Ernst Mayr and Bernhard

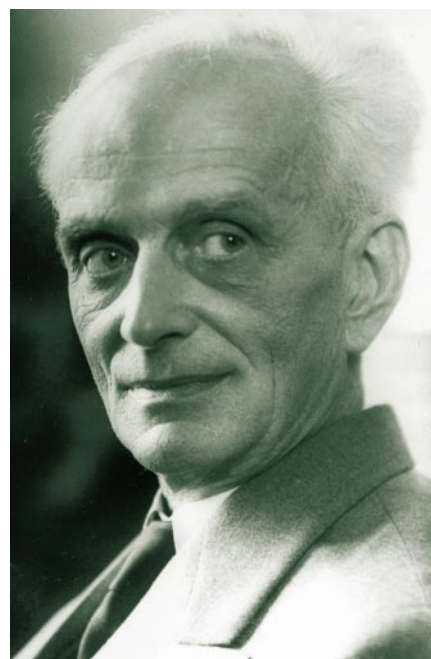
Rensch, were both guided by one man — Erwin Stresemann — who was pre-eminent in establishing many of the principles on which modern ideas of species are based.

Given Stresemann's paramount contribution to transforming the old typological and stagnant concept of species into the dynamic concept of biospecies (as one of many currently discussed species concepts), it is surprising that little information was previously available on his life and scientific work. The gap is now filled by this well-researched account of both the scientific and personal aspects of Stresemann's long and outstanding career, by three authors who, independently and at different times, came to know him.

Klaus Wunderlich, who examined Stresemann's scientific legacy in the *Staatsbibliothek Preussischer Kulturbesitz* in Berlin, provides the biographical information, while Erich Rutschke shows Stresemann the man. Most of the book comprises a description and analysis of Stresemann's scientific work and achievements by Jürgen Haffer, an experienced taxonomist, ornithologist and historian of science. An appendix contains previously unpublished extracts of manuscripts, including an unfinished manuscript from 1952 on the historical development of systematic knowledge on birds of paradise, one of Stresemann's lifelong obsessions, and letters that supplement the correspondence of Stresemann with Mayr and other ornithologists that was edited by Haffer and published in *Ökologie der Vögel* in 1997.

Stresemann was not only an eminent bird systematist and one of the most outstanding ornithologists of the twentieth century, but also a charismatic biologist who was one of the founders of the 'new systematics' in zoology during the 1920s. He was educated at the University of Munich, and by the age of 24 had already made an expedition to the Moluccan Islands (1910–12), published several articles on their bird life and been invited to write the ornithological contribution to Willy Kükenhals's famous *Handbuch der Zoologie*. With his later masterpiece, the seminal and lasting volume of the *Aves* (1927–34), Stresemann initiated the transformation of avian taxonomy into a branch of modern biological science, thereby laying the basis for a 'new avian biology'.

Appointed a curator of the Berlin Natural History Museum of the Humboldt University in 1921, Stresemann established a school of young ornithologists. The Second World War abruptly ended this successful tradition in German natural sciences. The museum was left trapped behind the Iron Curtain in East Berlin, making it impossible for Stresemann to resume his leading position in ornithology. Nevertheless, in him the unity of German ornithology survived during this period, mitigating the separation of Eastern and Western scientists.



Stresemann transformed avian taxonomy. Birds of paradise (below) were a lifelong obsession.

As Mayr later generously acknowledged, he owed to Stresemann the idea of reproductive isolation as the defining property of species, which later proved most valuable for defining biological species in Mayr's 1942 and 1963 accounts. From 1914 onwards, Stresemann also discussed geographical isolation and small population sizes as underlying factors in speciation, thus favouring what later became known as the allopatric model of speciation. In the 1920s and 1930s Stresemann synthesized a balanced view of historical and ecological factors long before vicariance biogeography as the systematic analysis of distributional patterns became the dominant perception.

The book is an affectionate account of Stresemann's career. But it is not silent about his overly conservative and pessimistic attitude, from the 1950s on, resisting modernization in the delimitation of bird taxa and becoming theory-hostile to the nature of species and the systematic relations between bird orders.

To anyone interested in the development of scientific ideas, the book offers an absorbing insight into the remarkable biologist's life and his leading role in ornithology and bioscience when the latter was still a holistic discipline. Particularly, we should be indebted to Jürgen Haffer for his efforts to reveal Stresemann's long-forgotten contributions. Only when historians of science cease to make use of the rich records of influential scientists, as Stresemann certainly was, are those scientists truly dead. ■

Matthias Glaubrech is at the Institute of Systematic Zoology, Museum für Naturkunde of the Humboldt University, Invalidenstrasse 43, 10115 Berlin, Germany.



Floral tributes

Early science texts carried illustrations that are master-works in their own right.

H. W. Lack

For science, 1543 was a miraculous year. It saw the (posthumous) publication, in Nuremberg, of *De revolutionibus orbium coelestium libri VI*, Copernicus's revolutionary tract hypothesizing that the Earth might not be the centre of the Universe, as well as *De humani corporis fabrica* by Vesalius, the first modern textbook on human anatomy, which was published in Basel. Both works opened up new worlds, as did the German translation of Leonhart Fuchs's (1501–66) herbal, originally published in Basel a year earlier, *De historia stirpium*.

The works of Vesalius and Fuchs relied heavily on illustrations, and it was these woodcuts that effectively established a new field: scientific illustration. Whereas the texts by Vesalius and Fuchs are today of historical interest only, the images are not: the anatomical illustrations Jan Steven van Kalkar created for Vesalius in Padova, and equivalent works done by Heinrich Füllmaurer, Albrecht Meyer, Jerg Ziegler and others for Fuchs in Tübingen, are fully identifiable and rightly regarded as master-works in their fields.

There were, of course, precursors, for example Martin Schongauer's peony, Leonardo da Vinci's lily and Albrecht Dürer's iris. But these naturalistic illustrations remained unpublished for centuries, whereas those by van Kalkar, Füllmaurer and Meyer were integrated into texts and published. They immediately became known to the nascent scientific community as woodcuts and for generations were regarded as reference standards. Moreover, Fuchs's *De historia stirpium* was very quickly translated into German and published, as *New Kreüterbuch*, in 1543. Consequently, this work had an immediate impact on a large audience, not only on the restricted circle of the learned.

Fuchs's publications provide early pictorial evidence for the multitude of plants arriving from the Americas (among them

maize, pumpkin and French marigold), illustrating the beginning of that immense shift in world agriculture and horticulture usually called the Columbian Exchange. *De historia stirpium* also contains woodcuts of castor bean and bladder senna from Africa, and hemp, aubergine and peach from Asia. In marked contrast, Fuchs's text is an amalgam of facts observed with his own eyes and opinion expressed by authors of Greek antiquity, notably Dioscorides, and their commentators. There is a clear emphasis on medicinal plants, among them opium poppy, camomile and foxglove — the last a surprise, since its role in therapy for the weak heart was understood only much later. Just as Vesalius is said to have insisted on his students dissecting corpses, Fuchs is reported to have been one of the first professors to take his students on botanical excursions.

After the publication of his *magnum opus*, Fuchs continued to collect illustrations of the plant kingdom, in particular of plants from all parts of the world then known — tulips from Central Asia, Persian fritillary from the Near East, sweet orange from China, and tobacco, tomatoes and sunflowers from the New World, some of these being the first records in Europe.

Altogether, there is good reason why the quincentenary of Fuchs's birth is being celebrated. The University of Tübingen in Germany — where Fuchs was professor of medicine for almost 31 years, serving seven times as rector, chairing meetings between cardinals from Rome and religious reformers including Philipp Melanchthon, and playing a key role in university reform — has organized a series of lectures on his life and work, as well as three botanical excursions. The town and university are organizing an exhibition called "Leonhart Fuchs and his time", opening on 20 June. The Nonnenhaus, Fuchs's home in the centre of Tübingen, still exists and its garden is being restored.

Fuchs's manuscript notes and drawings finally ended up in Vienna. Here, the Austrian National Library will display the *Codex Fuchs* for the first time, as part of the exhibition "A Garden Eden" running from 15 May to 31 October 2001. This nine-volume work, which represents the sum total of Fuchs's botanical knowledge and opinion, containing 1,541 illustrations, can be compared to Leonardo's *Codex Atlanticus* in Milan.

Three major recent or forthcoming publications focus on Fuchs. Stanford University



Green gift: a rose from Fuchs's *magnum opus*.

Press has produced a monumental study by a group of American historians of science headed by Frederick Meyer, which is accompanied by a facsimile edition of *De historia stirpium*. Eugen Ulmer of Stuttgart is to publish an in-depth analysis of the *Codex Fuchs* by the Baumann family, and Taschen of Cologne will publish shortly, in an English/German dual language edition, another reprint of Fuchs's *De historia stirpium*.

Visitors can see fine portraits of Fuchs in oils at the Württembergisches Landesmuseum in Stuttgart and the Art Institute in Chicago. But his most famous legacy enlivens many thousands of gardens whose owners may never have heard of the man: although he never saw these very popular plants, the genus *Fuchsia* was dedicated to Fuchs years after his death.

H. W. Lack is at the Institute of Biology, Free University Berlin, 14195 Berlin, and at the Botanic Garden and Botanical Museum Berlin-Dahlem, 14191 Berlin, Germany.

FURTHER READING

Meyer, F. G., Trueblood, E. E. & Heller, J. L. *The Great Herbal of Leonhart Fuchs* (Stanford Univ. Press, Stanford, 2001).
Baumann, B., Baumann, H. & Baumann-Schleihauf, S. *Die Kräuterbuch-Handschrift des Leonhart Fuchs* (Ulmer, Stuttgart, 2001).
Fuchs, L. *New Kreüterbuch mit Beiträgen von Klaus Dobat und Werner Dressendörfer* (Taschen, Köln, 2001).

Fuchs's publications are early pictorial evidence of the multitude of plants arriving from the Americas.

A light and a dark side

Thomas Elbert and Sabine Heim

The gigantic network that makes up the human cerebral cortex is dynamically and unceasingly reorganizing itself. This continuously changing entity is not easy to track, but we can glimpse the living brain by observing the organization within zones representing the first stages of sensory input from the outside world. Primary auditory, visual and somatosensory cortices mirror the spatial arrangement of the respective peripheral receptors: tonal frequency (place in the cochlea), visual space (place on the retina) or body surface are represented in the form of maps imprinted on the cortical sheet.

Although genetically encoded programs control the connections of these maps from the periphery to the cortical destination, their organization ultimately depends on the efficacy of the synapses connecting the nerve cells comprising the network, which is affected by environmental experience. For instance, two receptors of the same fingertip are more frequently activated in sync than are two receptors of different digits. According to the hebbian learning model, synchronous stimulation would lead to connections between the representations of the same fingertip but to a separation from those of the other digits: representational zones are shaped by the temporal pattern of coincident experience. An alteration in behaviourally relevant input towards the cortex will trigger a reorganization — alteration — of the map. The representation of a fingertip, for instance, can be enlarged; representations of adjacent fingers can invade its territory; or the representation

of two fingers can become fused.

Musicians provided one of the first examples of this phenomenon in humans: string players usually practise for hours a day over many years. While playing, the digits of the left hand are continuously fingering the strings, but the right-hand task of manipulating the bow involves much less processing of information from individual fingers. In string players, the cortical representations of the somatosensory left-hand regions were found to be expanded compared with those of the right hand and compared with the left hand of non-musicians.

What holds for the somatosensory cortex also holds for auditory processing: larger responses are elicited in the auditory cortex when musicians hear their own instrument played than are elicited by tones from other instruments. Perceptual correlates of map alteration indicate superior performance — one 'bright side' of cortical plasticity.

The most remarkable alteration of cortical maps has been observed in amputees. The nonstimulated area — for instance the arm region — remains actively engaged in information processing of nearby zones. As a consequence, the adjacent face and shoulder areas invade what was formerly arm territory. There is a close association between this invasion and adverse symptoms such as phantom-limb pain and, in the auditory system, tinnitus. Because treatment designed to reverse this invasion reduces phantom-limb pain, the adverse response is probably an unwanted consequence of the dynamic brain. This maladaptive disaster reveals the dark side of cortical reorganization.

The functional organization of one level of the cortex is governed by the interplay of earlier and later representational stages of the sensory processing stream. In mammals, the thalamus (one of the highest processing stages of the reptilian brain) feeds sensory information into the cortex. The cortex in turn feeds processed information back to the thalamus. When these top-down projections are inactivated, thalamic functional organization is dramatically degraded. There are nearly ten times as many fibres projecting top-down from the representational cortex to the thalamus as there are bottom-up from the thalamus to the cortex. Through this route, top-down processing determines the way in which sensory information is organized, filtered and processed. What we perceive is not only a matter of the incoming event but also a highly edited configuration of top-down organization and reorganization.

It is not surprising, then, that lesions in the central nervous system can lead to alterations in functional organization at all

Cortical reorganization

"The brain's representations of the periphery are dynamic and continuously modified by experience."

processing levels. After a brain lesion there are initial deficits in behaviour, perceptual or cognitive skills, but often spontaneous relief of symptoms. Cortical reorganization may be crucial for such recovery. New intervention strategies in conditions such as hemiplegia, writer's cramp, phantom-limb pain, aphasia or language-based learning disorders provide a 'bright side' ready to be exploited by neurorehabilitation.

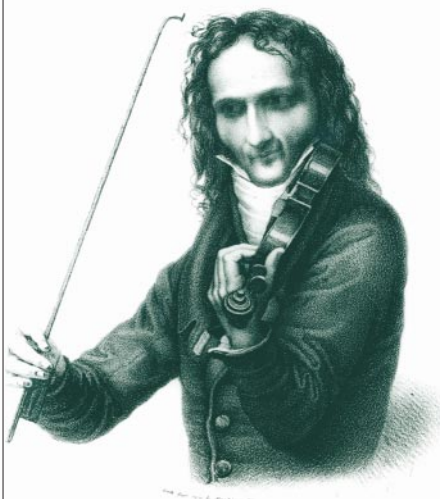
The dark side occurs when a peripheral lesion, by itself manageable, is triggered by an unusually intense, stressful experience to cause a negative or catastrophic cortical reorganization, such as that underlying phantom-limb pain and tinnitus. Fusion of representational zones is at the core of dystonia (such as musician's cramp or writer's cramp); Michael Merzenich's notion that dyslexia and even autism may be a consequence of inadequate functional organization (of the phonemotopic representation) stirs an exciting new avenue of research. And what about extreme experiences, such as traumatic stress or childhood abuse? Do they have the power to trigger reorganization of maps beyond the representational cortex, with consequent secondary disasters?

In a nonlinear, dynamic, self-organizing system such as the brain, small causes can have large effects. The brain should not be studied just in terms of monocular genetic and environmental influences. Learning is the adaptation of brain dynamics — when it has been disastrous, behavioural techniques (in conjunction with appropriate medication) may be the remedy. This prospect is the bright side of the dark side. ■

Thomas Elbert and Sabine Heim are in the Department of Psychology, University of Konstanz, Box D25, 78457 Konstanz, Germany.

FURTHER READING

- Buonomano, D. V. & Merzenich, M. M. *Annu. Rev. Neurosci.* **21**, 149–186 (1998).
- Gilbert, C. D. *Physiol. Rev.* **78**, 467–485 (1998).
- van Praag, H., Kempermann, G. & Gage, F. H. *Nature Rev. Neurosci.* **1**, 191–198 (2000).
- Elbert, T., Heim, S. & Rockstroh, B. Neural plasticity and development. In *Handbook of Developmental Cognitive Neuroscience*. (eds Nelson, C. & Luciana, M.) (MIT Press, Cambridge, MA, 2001).
- Nudo, R. J. *Curr. Opin. Neurobiol.* **9**, 740–747 (1999).



Cortical representation in musicians, such as Niccolò Paganini, reflects their art.

Evolving ideas of brain evolution

Jon H. Kaas and Christine E. Collins

Recent analyses of an old data set are starting to reveal patterns in the evolution of mammalian brains. The latest study shows that mammalian groups are characterized by basic similarities in brain proportions.

The brains of mammals vary greatly in size, shape, internal organization and functional capabilities. The human brain is much larger than that of a mouse lemur, for example (Fig. 1), and includes subdivisions and connections not found in a lemur. Those of us who are interested in how these differences evolved have only a few sources of information. A useful, if fragmentary, historical record is available from 'endocasts' of fossilized skulls: occasionally, during fossilization, the space inside the skull fills with silt which then hardens, producing a cast of the internal structures. This record can be very accurate, but it mainly shows us the sizes and shapes of brains and the locations of large brain features such as fissures (grooves). It provides few clues to internal organization. So theories about brain evolution have also depended on comparative studies of the brains of living mammals, and Clark, Mitra and Wang¹ describe one such analysis on page 189 of this issue.

The general premise of comparative studies is that physical features that are found in several species have been retained from their common ancestor, and that features that are shared only by closely related species must have evolved recently. The challenge for those using this approach is to get reliable and comparable observations from the brains of a suitable range of living species. So it is perhaps not surprising that, although it is 20 years old, one of the most extensive studies² of the sizes of brain parts in mammals still forms the basis for many comparative analyses of mammalian brain evolution. In this paper², Stephan *et al.* published the volumes of easily identified brain structures for an impressive number of insectivore, bat and primate species.

In one analysis³ of these data, Finlay and Darlington concluded that most of the differences among mammalian brains result from the disproportionately greater growth in larger brains of the later-maturing structures of the forebrain. Their theory emphasized the overriding importance of a basic plan for brain development that had been modified during evolution according to the size of the brain. More recently, Barton and Harvey⁴ reconsidered the data² and argued that, in general, the sizes of brain parts evolve independently of each other, although some functionally related parts grow larger together. In primates, such related brain

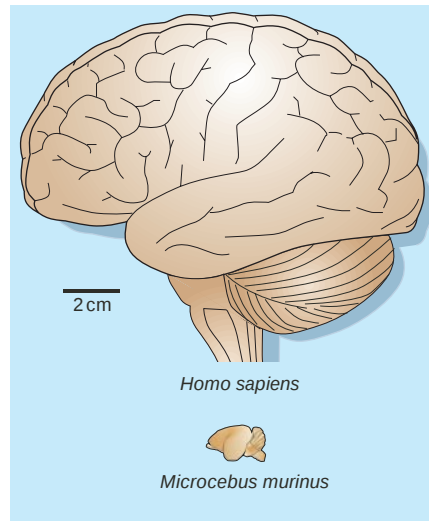


Figure 1 Side-on views of the brains of a human and a small prosimian primate, the mouse lemur (*Microcebus murinus*). Although these brains are very different in size, their proportions are similar, and group together with those of other primates, as Clark *et al.* show¹.

parts include the cerebellum (involved in processing sensory information to coordinate movements) and the neocortex (also involved in processing sensory information). Soon afterwards, de Winter and Oxnard⁵ used multivariate statistical methods to provide evidence for other patterns of brain evolution, including the evolution of similar brain proportions in distantly related species — such as woolly monkeys and apes — that behave similarly.

Clark *et al.*¹ have now compared the sizes of different brain parts to total brain size, by using multidimensional comparisons of brain proportions. They define what they call 'cerebrotypes' for the species they have examined. So, for example, in tree shrews the fraction of the brain occupied by the telencephalon is about 61%, and that taken up by the cerebellum is about 13%. The combination of such proportions makes up the cerebrotypes for this species. One of the authors' findings¹ is that different taxonomic groups have distinct cerebrotypes. So insectivores, which are now thought to have evolved from several ancestors rather than just one⁶, contain separate taxonomic groups, and these groups have different cerebrotypes. In addition, the cerebrotypes of tree shrews clearly do not group with those of

primates, fitting with the fact that tree shrews are no longer considered to be primates⁶.

Furthermore, the cerebrotypes for primate groups can be divided into largely separate subtypes for prosimians (such as lemurs), New World monkeys, Old World monkeys and hominoids (apes and humans). The overlap in the cerebrotypes of some species of Old and New World monkeys suggests that these species have similar lifestyles, and that their brains have therefore evolved in similar ways. Overall, the presumed length of time since diverging from a common ancestor correlates well with differences in cerebrotypes for primate groups. Observations of the proportional size of the cerebellum also lead to the interesting conclusion that its size, and therefore functions, relate broadly to the rest of the brain. However, in animals with certain specialized functions — such as echolocation in dolphins, whales and small bats — the cerebellum is disproportionately large compared with that in other species.

The differing conclusions that have been reached^{1,3-5} from analysis of the same data² are not necessarily contradictory, as they reflect different approaches and emphasize different aspects of the data set. Indeed, the outlines of patterns of brain evolution are starting to emerge. It seems fair to conclude that mammals have a characteristic pattern of brain development, which has been distorted during evolution as brains became larger. It also seems reasonable that the size of one brain structure in relation to another varies from species to species, and that these variations occur in ways that should reflect the importance of the function of that structure. Finally, groups that share a common ancestor are characterized by basic similarities in brain organization.

But we are far from a complete understanding of brain evolution. We need to know more about the details of brain organization (the sizes, numbers, identities and interconnections of subcortical nuclei and cortical areas) for more species, and what these differences might mean for function. We already know, for example, that two mammalian groups — carnivores and primates — have different patterns of layers in part of the visual thalamus, but we don't fully understand how these two patterns relate to visual abilities.

In addition, we might productively con-

sider theoretical problems and solutions to the evolutionary challenge of making brains bigger or smaller⁷. For other body organs, as well as for bridges and buildings, the problems of changes in scale have been investigated thoroughly⁸. It is now time to do the same for brains and to find out whether theory and evolution would cope with these problems in similar ways. ■

Jon H. Kaas and Christine E. Collins are in the Department of Psychology, Vanderbilt University, Nashville, Tennessee 37203, USA.

e-mails: jon.kaas@vanderbilt.edu

christine.collins@vanderbilt.edu

1. Clark, D. A., Mitra, P. P. & Wang, S. S.-H. *Nature* **411**, 189–193 (2001).
2. Stephan, H., Frahm, H. & Baron, G. *Folia Primatologica* **35**, 1–29 (1981).
3. Finlay, B. L. & Darlington, R. B. *Science* **268**, 1578–1584 (1995).
4. Barton, R. A. & Harvey, P. H. *Nature* **405**, 1055–1058 (2000).
5. de Winter, W. & Oxnard, C. E. *Nature* **409**, 710–714 (2001).
6. Springer, M. S. & de Jong, W. W. *Science* **291**, 1709–1711 (2001).
7. Kaas, J. H. *Brain Mind* **1**, 7–23 (2000).
8. Schmidt-Nielsen, K. *Scaling: Why is Animal Size So Important?* (Cambridge Univ. Press, New York, 1984).

Climate change

The Indonesian valve

James D. Wright

The behaviour of the North Atlantic is often invoked to explain the effects of climate change. But for certain episodes, including perhaps a period in human evolution, events elsewhere may have had a greater influence.

The climate of East Africa became drier between about 5 million and 2.5 million years ago, and that may have been the catalyst that forced our ancestors to adapt to a savannah environment as the forests dwindled^{1,2}. At about the same time, the Earth entered a climate mode dominated by the waxing and waning of large continental ice sheets. The coincidental timing of global cooling, African aridity and human evolution invites speculation about a common link³. For that, we must look to the oceans — in redistributing heat and influencing greenhouse-gas concentrations globally, they are the main component in determining climate change. Marine records tell us that the transition to large-scale glacial cycles took at least a million years; and plate-tectonic motions that opened or closed ocean gateways are thought to have triggered these events⁴.

On page 157 of this issue, Cane and Molnar⁵ present an analysis of changes in surface-ocean circulation that they believe occurred as an oceanic gateway — the Indonesian seaway — narrowed over the past 5 million years. This gateway acts as a valve for water flowing from the Pacific into the Indian Ocean. Plate tectonics in the Indonesian region is complicated, but Cane and Molnar show that the passages regulating water flow from the Pacific to the Indian Ocean 5 million years ago were wider and deeper, and were located further to the south, than they are today. Surface water in the South Pacific is warmer and saltier than that in the North Pacific. Cane and Molnar argue that the more southerly position for the Indo-Pacific connection meant that the warmer South Pacific flowed into the Indian Ocean. The result was warmer sea surface temperatures in the Indian Ocean and high levels of evaporation and precipitation — and wet East African climates.

Over the past 5 million years, the constriction and northern movement of the Indonesian seaway have progressively shut off the South Pacific source of water, while increasing the influence of the colder North Pacific. These changes should have cooled the tropical Indian Ocean and reduced the precipitation, leading to a gradual drying of East Africa. Again, then, the idea is that the Indonesian seaway, controlled by the northern movement of New Guinea and smaller islands, has acted like a valve, regulating the relative amount of warm and cool water entering the Indian Ocean.

A more speculative aspect of Cane and Molnar's paper deals with the possible effects on global climate of this narrowing of the Indonesian seaway. The authors argue that, when the seaway was farther south, conditions in the tropical Pacific would have been more like those observed during modern El Niños (that is, both east–west and vertical thermal gradients in the ocean would have been weaker). This configuration would have promoted greater heat transport to the high northern latitudes than at present; and that higher heat flux would have inhibited the growth of large ice sheets in the Northern Hemisphere. In support of Cane and Molnar's speculation, palaeoceanographic reconstruction of the tropical Pacific between 5 million and 3 million years ago matches the prediction of smaller east–west and vertical temperature differences⁶.

This speculation is especially provocative because it requires a new principle for understanding the glacial cycles that developed 3.5–2.5 million years ago. Existing models for large-scale Northern Hemisphere glaciation focus on increased circulation of the North Atlantic 'conveyor', which includes the Gulf Stream, as the cause of ice-sheet development^{4,7–9}. The finer points in these

models vary, but most of them begin with increased precipitation at high northern latitudes as a result of a more vigorous Gulf Stream. And going back one step, the closure of the Panamanian isthmus is seen as the trigger for initiating the more powerful Gulf Stream.

Recently, however, a fly in the ointment for these hypotheses has appeared¹⁰. An evaluation of North Atlantic circulation for between 5 million and 2 million years ago contradicts the enhanced Gulf Stream mechanism for Northern Hemisphere glaciation. The new data¹⁰ indicate that the North Atlantic conveyor became considerably weaker, not stronger, 3.5–2.5 million years ago.

Cane and Molnar⁵ propose a shift in thinking away from the North Atlantic to the Indonesian gateway as a factor governing global climate change. There are good reasons to take this idea seriously, speculative though it may be, because the Indonesian region is undoubtedly important in the redistribution of heat received at the Earth's surface and in moisture fluxes to the atmosphere. Moreover, Cane and Molnar make several predictions that can be tested.

The most obvious prediction is that sea surface temperatures in the Indian Ocean cooled between 5 million and 2 million years ago. This may be difficult to test, however. The change in temperature was probably 2–3 °C. Oxygen-isotope analysis of planktonic foraminifera, a useful tool for estimating past sea surface temperatures, can resolve this change. But because of evaporation, the South Pacific is saltier, as well as warmer, than the North Pacific. The evaporation that increases the salinity also increases the oxygen-isotope value of the surface water, offsetting the temperature effect. Advances in measuring the Mg/Ca ratios in foraminifera, and then using those ratios to estimate temperature, may solve the problem.

My own view is that Cane and Molnar are correct in their view that African aridity is linked to sea surface temperatures in the Indian Ocean, and that the most likely cause of cooling there was a narrowing of the Indonesian seaway. I am less confident that these changes had much to do with glaciation of the Northern Hemisphere, for one simple reason: from 10 million to 5.6 million years ago, cyclic glaciation was highly active in the Northern hemisphere and glaciation was suppressed between 5.5 million and 3.5 million years ago. Moreover, changes in the North Atlantic conveyor circulation cannot be ruled out in driving these glaciations. The conveyor delivers a substantial amount of heat to the high northern latitudes; the link with glaciation might have been through reduced heat fluxes as conveyor circulation decreased¹¹, rather than through precipitation. Nonetheless, Cane and Mol-

nar have armed the palaeoclimate community with the theory and predictions that will allow us to examine events in the Pacific, Indian and Atlantic oceans — especially the valves controlling water flow into and out of them — as drivers of Earth's climate. ■

James D. Wright is in the Department of Geological Sciences, Rutgers, The State University of New Jersey, 610 Taylor Road, Piscataway, New Jersey 08854-8066, USA.

e-mail: jdwright@email.rci.rutgers.edu

- Behrensmeier, A. K., Todd, N. E., Potts, R. & McBrinn, G. *Science* **278**, 1589–1594 (1998).

2. Vrba, E. S. in *Evolutionary History of the "Robust" Australopithecine* (ed. Grine, F.) 405–426 (de Gruyter, New York, 1988).
3. deMenocal, P. B. *Science* **270**, 53–59 (1995).
4. Haug, G. H. & Tiedemann, R. *Nature* **383**, 673–676 (1998).
5. Cane, M. A. & Molnar, P. *Nature* **411**, 157–162 (2001).
6. Chaisson, W. P. & Ravelo, A. C. *Paleoceanography* **15**, 497–505 (2000).
7. Keigwin, L. D. Jr *Earth Planet. Sci. Lett.* **45**, 361–382 (1979).
8. Driscoll, N. W. & Haug, G. H. *Science* **282**, 436–438 (1998).
9. Stanley, S. M. J. *Paleontol.* **69**, 999–1007 (1995).
10. Ravelo, A. C. & Andreasen, D. H. *Geophys. Res. Lett.* **27**, 1001–1004 (2000).
11. Berger, W. H. & Wefer, G. in *The South Atlantic: Present and Past Circulation* (ed. Wefer, G. et al.) 363–410 (Springer, Berlin, 1996).

Evolution

Developmental circuits rewired

Eörs Szathmáry

Body segmentation occurs during the development of many invertebrate animals. Advances are being made by those striving to produce computer models of the genetic networks underlying the process.

You might think that key features of embryo development would not be easily altered during evolution. Such changes should be lethal, but they are surprisingly common. So how can the underlying genetic regulatory circuits be rewired without making the whole population of organisms inviable? Writing in the new journal *Evolution and Development*^{1,2}, Salazar-Ciudad *et al.* show that evolution favours alternative genetic circuits, depending on the size and the complexity of the patterning task to be accomplished.

It has long been known that the beginning and the end of embryological development are fairly variable traits in evolution. Diversity at an early developmental stage (in the mechanism of gastrulation, for example) can be attributed to evolutionary adaptations to the ecological setting in which the embryo begins to unfold. The later phases must also differ, otherwise species would all look the same³.

Development in parasitic wasps has diverged widely, for example. Early cell divisions of the fertilized egg, establishment of head-to-tail polarity, and the genetic circuit for body segmentation (the mechanism by which stripes are created through intergene regulation) have all been modified, apparently as adaptations to the parasitic lifestyle⁴. A wasp known as *Copidosoma*, for instance, has many features resembling those of mammalian development, apparently to enable the embryo to tap into the host for nutrition. But developmental malleability can go beyond this. Although the external appearance of an organism may be fixed, the genetic network — in which genes switch one another on and off so that programmed development runs successfully — seems to be changeable. This is analogous

to rewiring your laptop without changing the housing. It seems to have occurred in the fruitfly *Drosophila*, in which two genes, *fushi tarazu* and *bicoid*, appear to have been recruited for segmentation as an afterthought⁵. At first sight, the observation seems to run counter to the finding⁶ from computer modelling that the genetic network responsible for setting up polarity within segments is highly robust to change. So, how does one reconcile conservation and change?

As they describe in their first paper¹, Salazar-Ciudad *et al.* have simplified the question by concentrating on a row of cells (or cell nuclei where there is initially no membrane between them, as in *Drosophila*). It is in this row of cells that genes interact by regulating the activities of some other genes in the system (Fig. 1). A series of artificial selection experiments is carried out in the computer until a pre-specified goal pattern is reached. Perhaps the most enchanting feature of this model is that it puts embryonic pattern formation, and the evolution of pattern formation, into a unified theoretical framework.

One of the pre-specified tasks is segmentation. Similar phenomena, including stripes on animal coats, fascinated Alan Turing, who in 1952 proposed a mechanism for pattern formation⁷. He showed that, in a chemical system that starts off as spatially homogeneous, a diffusing activator and an inhibitor could give rise to stationary-wave-like concentration profiles of chemicals (Fig. 2a). Similar reaction–diffusion mechanisms may be at work in some biological systems, such as in the skin of the angelfish⁸. But other systems, previously thought to be produced by a Turing-like system, such as stripe formation in *Drosophila*, seem to use a mechanism⁹

in which each stripe's identity is determined by a different combination of regulatory elements, following an initial spatial heterogeneity (Fig. 2b). The underlying genetic circuit has a hierarchical structure³.

Several conclusions emerge from Salazar-Ciudad and colleagues' model¹. First, evolving networks always fall into one of the two categories: 'emergent' (reaction–diffusion) or hierarchical¹⁰. Second, the network producing the target pattern tends to be hierarchical when the number of stripes to be generated is fewer than four, and to be emergent above four; also, emergent networks generally produce the more complex patterns. Third, emergent networks are more changeable than hierarchical ones when subject to mutations — this means that the patterns become more dissimilar from the original one than in the case of hierarchies (deviation from the morphological norm can often be selectively disadvantageous). Fourth, hierarchical networks can produce more finely tuned patterns than emergent ones.

Together, these results indicate that a whole set of possible hierarchical networks covers a larger territory in 'morphological space' (understood as the set of possible gene-expression patterns in the modelled row of cells) than a whole set of emergent

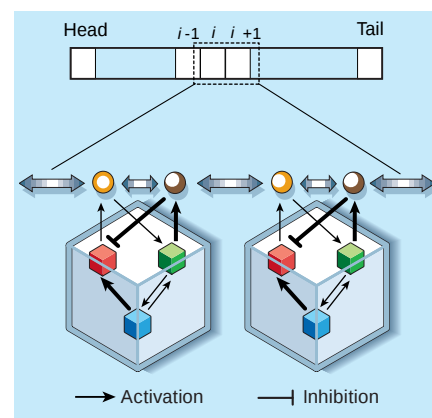


Figure 1 A model for the evolution of pattern generation. Salazar-Ciudad *et al.*¹ model the early developing embryo from head to tail as a row of 25 cells (or cell nuclei) in which genes (boxes) and diffusible factors (circles) interact. Two cells, occupying the *i*th and (*i* + 1)th position in the row, are enlarged here. Each cell has the same set of genes and the same intergene interactions. The network functions as a nonlinear dynamical system. The concentrations of non-diffusible and diffusible gene products are calculated numerically, and a pattern is generated when at least one gene product attains different levels in different cells. (Modified from ref. 2.) Overall network evolution is simulated with a population of 100 such networks, the 'fitness' (viability) of each being inversely proportional to the distance of its generated pattern from the target pattern. Changing the interaction weights randomly generates the variation on which selection acts.

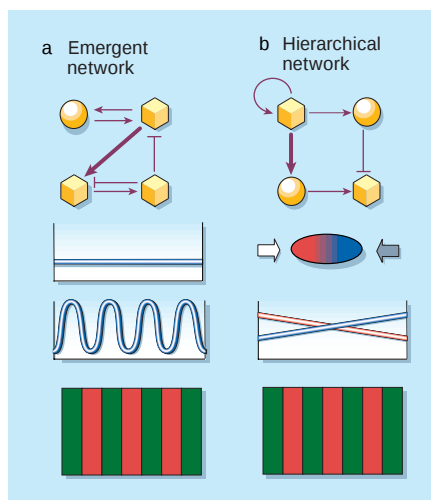


Figure 2 Two routes to segmentation. a, Emergent networks produce a standing wave of a signal molecule (morphogen) that acts as a cue for segment formation; the identity of each segment is determined by a critical level of the same morphogen. These networks are also known as reaction–diffusion or Turing-like. b, Hierarchical networks produce the same pattern because each stripe depends on a unique combination of morphogen concentrations. This is the mechanism that applies to *Drosophila*, the necessary initial heterogeneity in the developing egg being produced by the mother. The networks at the top are examples of the emergent and hierarchical types.

networks. This implies that a greater diversity of morphologically different — but still similar — species could be produced by variation of, and selection on, hierarchical gene regulatory networks.

It is also remarkable that when an emergent and a hierarchical network produce the same complex pattern, the emergent network requires fewer genes to do it. The reverse is true for simple patterns. Finally, the authors argue that, once established, emergent networks producing complex patterns are likely to be gradually overwritten, as on a palimpsest, by a hierarchical one.

In the second paper², the authors use these conclusions to inform their analysis of the evolution of segmentation. Segmentation seems to be an ancestral characteristic of arthropods, including spiders¹¹, but there is a lot of variation. In *Drosophila*, for instance, there are no membranes between nuclei in the critical period (the structure is called a syncytium), and the segments arise simultaneously. In other organisms, such as the flour beetle *Schistocerca*, membrane-enclosed cells are involved, and segments are produced sequentially from a posterior zone. Intermediate types also exist³. Newman¹² has hypothesized that the evolutionary transition from the sequential–cellular to the simultaneous–syncytial modes of segmentation could have occurred if the former was produced by a network-controlled

clock mechanism coupled to directional cell proliferation. The new work² shows that emergent networks with clock-like behaviour can often produce segments when the boundaries between cells are ‘knocked out’, corresponding to the emergence of the syncytium.

So an evolutionary shift from cellular segmentation to the non-cellular type found in *Drosophila* could have occurred when a circuit of the emergent type with clock-like behaviour continued to act in a syncytium. The emergent-type network has presumably been replaced later by a more elaborate and fine-tuned hierarchical one. This transition has not yet been the subject of much modelling work. But some simulations indicate that hierarchical networks can indeed replace emergent ones, and that the hierarchical networks are more likely to give rise to conserved modules. There is plenty more to come from deciphering

these developmental palimpsests from the powerful combination of studies *in silico* and *in vivo*.

Eörs Szathmáry is in the Department of Plant Taxonomy and Ecology, Eötvös University, and the Collegium Budapest (Institute for Advanced Study), 2 Szentháromság, H-1014 Budapest, Hungary.

e-mail: szathmary@colbud.hu

1. Salazar-Ciudad, I., Newman, S. A. & Solé, R. V. *Evol. Dev.* **3**, 84–94 (2001).
2. Salazar-Ciudad, I., Solé, R. V. & Newman, S. A. *Evol. Dev.* **3**, 95–103 (2001).
3. Wolpert, L. et al. *Principles of Development* (Current Biology/Oxford Univ. Press, 1998).
4. Grbic, M. *BioEssays* **22**, 920–932 (2000).
5. Gibson, G. *Curr. Biol.* **10**, R452–R455 (2000).
6. Von Dassow, G., Meier, E., Munro, E. M. & Odell, G. M. *Nature* **406**, 188–192 (2000).
7. Turing, A. *Phil. Trans R. Soc. Lond. B* **237**, 37–72 (1952).
8. Kondo, S. & Asai, R. *Nature* **376**, 765–768 (1995).
9. Akam, M. *Nature* **341**, 282–283 (1989).
10. Salazar-Ciudad, I., Garcia-Fernandez, J. & Solé, R. V. *J. Theor. Biol.* **205**, 587–603 (2000).
11. Damen, W. G. M., Weller, M. & Tautz, D. *Proc. Natl Acad. Sci. USA* **97**, 4515–4519 (2000).
12. Newman, S. A. *BioEssays* **15**, 277–283 (1993).

Astronomy

Young stars go cruising by

Gibor Basri

Astronomers have spotted a large number of newly formed stars near our Solar System. So there might be more going on in our Galactic neighbourhood than we thought.

A new perspective on the history and contents of our little section of the Galaxy is emerging from data collected by various satellites. Our Solar System appears to be immersed in a spray of young stars from distant star-forming regions. Nearly 100 astronomers (many of them also fairly young) who are working to identify young stars near Earth gathered at the NASA Ames Research Center for the first ever workshop* on this new subfield. These stars may provide us with the best chance of directly imaging extrasolar planets sooner rather than later.

Regions of star formation are marked by dense interstellar clouds of molecular hydrogen with about 1% by mass of ‘dust’ (tiny grains of silicate and carbon-based minerals) that blocks our view through them. When hot, young stars form in these clouds, they produce regions of ionized hydrogen that glow and are easy to find. These regions contain many young stars (20–50 million years old), often still surrounded by circumstellar disks of gas and dust, which could be sites of planet formation. The closest star-forming regions are about 400 light years away, and the nearest giant molecular cloud (in Orion) is about 1,000 light years away — for comparison, our Galaxy is 100,000 light years across.

In the past few years, several satellites have provided us with a more complete picture of the young stars in our neighbourhood. An all-sky survey by the ROSAT X-ray observatory (flown by the European Space Agency) revealed all the strong X-ray emitters near the Sun. Young stars have vigorous magnetic activity, which means that their X-ray emission can be up to a thousand times brighter than our Sun’s. The Hipparcos satellite (also European) has given astronomers more accurate positions for many nearby stars. This gives a better idea of the distances to many stars through the use of trigonometric parallax, and a much more accurate measure of their ‘proper motion’ (the sideways motion of stars in the sky, relative to the distant background).

The X-rays recorded by ROSAT indicate whether a star might be young, and ground-based telescopes can then be used to obtain spectra and test for the presence of lithium. Lithium is a good indicator of a young star, as it is destroyed by nuclear fusion before a star is 20–50 million years old. Spectra also provide the line-of-sight velocity of the star (through the Doppler shift), which can be combined with the proper motion to get a three-dimensional velocity vector (space motion) for a star travelling through the local neighbourhood. Finally, the distances measured by Hipparcos can be used to convert a star’s observed brightness to an

*Young Stars Near Earth Workshop. NASA Ames Research Center, Mountain View, California, USA. 28–30 March 2001.

intrinsic luminosity and therefore determine whether the star is mature enough to be a stable main sequence star given its temperature.

The results presented at the Young Stars Near Earth workshop suggest that the Solar neighbourhood is infused with young stars, many of them travelling in loosely bound groups. This indicates that they formed together in small molecular clouds, which disappeared long ago — and now the stars are dispersing as well. A number of them can be traced back (given that they formed perhaps 30 million years ago) to the vicinity of what is still a very active star-formation complex in the general direction of the constellations Scorpius (Sco) and Centaurus (Cen).

Perhaps the best known of these groups is called the TW Hya association. It takes its name from the star TW Hydra, which has been known for a couple of decades to be a T Tauri star — one which still has an actively accreting circumstellar disk. It was always a mystery what it was doing only 150 light years from us, nowhere near any molecular clouds. About 20 other stars have since been found to have the same space motion as TW Hya, spread over a volume perhaps 50 light years across. Similar associations have been found in the directions of the constellations Chameleon, Horologium and Tucana. It is unclear to what extent these associations are really distinct from each other, or are part of a more general flow of young stars from the direction of Sco–Cen. One complaint at the meeting was that all the good targets seem to be visible only from the Southern Hemisphere; most astronomers are in the Northern Hemisphere, although times are changing.

A few nearby bright stars that still have remnant dust disks now appear to be part of the young local population. These Vega-type stars tend to be so hot that none of our traditional tests for youth work on them, but their space motions and association with other young stars suggest that they too are relatively young. NASA hopes to fly the Full-Sky Astrometric Mission (FAME) mission in about three years, which will have much greater accuracy and sensitivity than the Hipparcos satellite, and will measure the whole sky. The FAME mission would be an enormous boon to this field; it should allow us to understand the motions of most Solar-type stars within a few hundred light years of the Sun.

The study of very young stars has much to offer. They are relatively bright, which allows us to obtain more detailed spectra and study them more closely. As we build bigger telescopes we should be able to study in detail the disks (active or remnant) that we know exist around some nearby stars. Imaging studies of nearby young stars are also in the best position to detect substellar companions (brown dwarfs and giant planets) of these

stars. Substellar objects fade with time, but can be relatively bright when they are very young. Astronomers are now looking as closely as possible at all the newly discovered stars with better-resolution and higher-contrast observations.

It will not be surprising if the first real image of an extrasolar giant planet is obtained in one of these systems. Indeed, there are already good candidates. An object in the TW Hya association would definitely be an extrasolar planet if it really is a companion to the young star. This question can be resolved by studying their proper motion, as a faint background star (rather than a planet) will not join in the motion of the nearby star across the sky. A team using adaptive optics at the Keck telescope in Hawaii had hoped to announce the answer at the workshop, but their data were still too ambiguous. Even if this object does not pan out, these studies could eventually provide much more information on extrasolar planets than just the mass and orbit given by current techniques.

The overall picture that emerges from the

workshop is that there has been vigorous star formation in the Sco–Cen region for perhaps the past 100 million years. The main action is more than 1,000 light years away. We know that the Sun sits in a hot evacuated bubble in the interstellar medium, perhaps 1,500 light years across. This was created when a number of massive stars exploded as supernovae and cleared out and heated this bubble. This violence may have compressed clouds that were already moving in our direction, triggering star formation in them. The process of star formation destroys a molecular cloud in about 10 million years, and the stars that formed will then typically disperse. It is these stars that we are now seeing in the young nearby population. They have travelled far to meet up with us and yet we have only recently arrived at this location ourselves — we were more than 1,000 light years away when our new neighbours were born. ■

Gibor Basri is in the Astronomy Department, University of California at Berkeley, California 94720-3411, USA.

e-mail: basri@astro.berkeley.edu

Genomic stability

Hip-hopping out of control

David E. Symer and Judith Bender

‘Jumping genes’ can wreak havoc by hopping about a genome. Some plants can keep them under control by modifying them with certain chemical groups.

Over the past few years there has been heightened interest in understanding the ramifications of ‘epigenetic’ changes to genomes. Such alterations are heritable, non-random and intricate, and occur in nearly all organisms. But they do not directly change DNA sequences. Instead, epigenetic processes include the reversible attachment of certain chemical groups to genomic DNA or to the proteins that package it into a tidy, compact form known as chromatin.

These modifications are essential for gene expression and silencing, and the inactivation of one of the two X chromosomes in female mammals. They are also involved in the development of cancer, and possibly ageing. And, as highlighted on page 212 of this issue¹ and in *Genes and Development*², they are also required for the regulation of plant transposable elements — the ‘jumping genes’ discovered by Barbara McClintock. Miura and co-workers¹ and Singer and colleagues² have dramatic evidence that defects in epigenetic modifications lead to transposable elements (transposons) hopping about the genome of thale cress, *Arabidopsis thaliana*, much more often than usual.

In essence, transposons are discrete segments of DNA that can move from one site

in the genome to another within a cell. Genome-sequencing projects have revealed that these mobile elements are likely to occur at a high frequency, and in many forms, in eukaryotic (non-bacterial) organisms, including *Arabidopsis*³.

Transposons are typically covered with methyl groups, so it has been proposed that this epigenetic modification evolved, at least in part, as a defence mechanism to limit the spread of such ‘junk’ DNA⁴. In the delicate balance between the damage that parasitic elements such as these can wreak in their quest to proliferate, and the occasional benefit that transposon movement may afford the host, methylation would provide a flexible means of regulating transposon jumping. Indeed, long-term studies of maize have shown that transposon activity inversely correlates with methylation, which can be altered during different stages of development^{5,6}. In general, methylation is likely to work by preventing genes from being expressed (transcribed). So, another advantage of methylating transposons would be to reduce the pointless expression of RNA (so-called transcriptional noise) from transposon-encoded genes, most of which are riddled with mutations and so cannot function⁷.

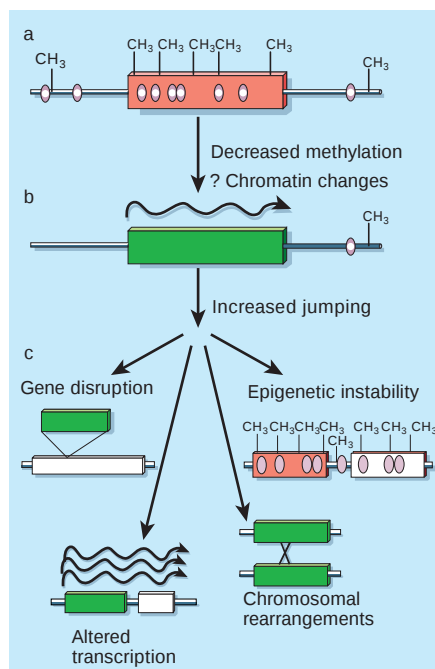


Figure 1 Possible causes and consequences of uncontrolled movement of transposable elements ('jumping genes'). Red and green boxes represent silenced and active transposons, respectively. **a**, A transposable element in the genome is typically silenced by methylation (addition of CH₃ groups) and probably by repressive chromatin structures — certain modifications of the proteins that bundle up the genome (purple ovals). **b**, When these controls are disrupted — for example, in plants with the *ddm1* mutation^{1,2} — the transposon is expressed (represented by a wavy line) and jumps about the genome at a much higher rate. **c**, Possible consequences of the transposon jumping into new sites elsewhere in the genome, such as into or nearby other genes (open rectangles).

The results of Miura *et al.*¹ and Singer *et al.*² provide striking support for the genome-defence model⁴. They show that at least two distinct kinds of transposon in *Arabidopsis* are activated and start jumping specifically in the context of reduced DNA methylation (Fig. 1). Both groups used *ddm1* (for 'decrease in DNA methylation') mutant plants, which have genome-wide decreases in methylation. The normal DDM1 protein is similar to some chromatin-remodelling proteins³, so the mutant plants may also have altered chromatin structures, another type of epigenetic modification.

Using a predictive bioinformatics approach, Singer *et al.*² identified a family of 22 transposons — related to maize transposons called Mutator elements — scattered throughout the *Arabidopsis* genome. At least one of these transposons readily hopped about the genome of *ddm1* mutants. By contrast, Miura *et al.*¹ started by investigating a morphological abnormality that arose spontaneously in an inbred lineage of *ddm1* plants. They showed that this defect was

caused by a transposon hopping into, and disrupting, a gene encoding a key protein involved in biosynthesis. The transposon was from the CACTA family, so called because the ends of these transposons have the DNA sequence CACTA. Miura *et al.* found three other related transposons in *Arabidopsis*, and at least two of the four jumped about avidly in *ddm1* plants.

Neither group found appreciable transposon movement in normal plants, indicating that methylation efficiently suppresses these mobile DNA elements. Moreover, a previous study found that the transcription of a different transposon is activated in *ddm1* plants⁹. Similarly, the expression of mouse transposons is activated in methylation-deficient animals^{10,11}. Collectively, these studies give considerable support to the genome-defence model⁴. Another implication is that loss of methylation of transposons may also result in increased transcriptional noise, regardless of the consequences for transposon movement.

Like any good experiment, these studies^{1,2} raise further questions, some specific and others broader. One is whether the *ddm1* mutation has effects on transposons — other than increased transcription — that might contribute to their increased mobility. For example, might it alter the chromatin structure at the transposons' donor or target sites? The biochemical functions of the normal DDM1 gene are not yet certain, so the basis for methylation changes in *ddm1* mutant plants remains unclear. Another question is whether mutations in other *Arabidopsis* genes that alter methylation or chromatin structure will have similar consequences, or perhaps affect only particular subsections of the genome or classes of mobile elements.

Certainly, the new work^{1,2} will allow the generation of a practical resource, consisting of transposons sprinkled across the genome. This will be useful for investigating the possible consequences of transposon mobilization (Fig. 1), including gene disruption¹, altered gene transcription, chromosomal rearrangements and epigenetic instability¹².

It remains to be seen whether other *Arabidopsis* transposons are mobilized in methylation-deficient mutants. And it will be interesting to find out whether transposons in animal cells are mobilized by disrupting methylation patterns or chromatin structures. If so, this finding might give pause to those contemplating treating patients with drugs that perturb genome-wide epigenetic processes, such as chemotherapeutic drugs intended to 'unsilence' aberrantly hyper-methylated and repressed genes in cancers.

These studies open new vistas on transposons and epigenetic controls, which together have shaped eukaryotic genomes through evolution, and provide a useful way of studying jumping genes. They also highlight the destabilizing effects of perturba-



100 YEARS AGO

The first sealed thermometer was made some time prior to 1654 by Ferdinand II., Grand Duke of Tuscany; he filled the bulb and part of the tube with alcohol, and then sealed the tube by melting the glass tip. Ferdinand and his brother, Leopold de Medici, promoted the establishment in Florence of the Accademia del Cimento, and the account of their experiments, published in 1667 and translated into English by Waller in 1684, contain descriptions of various thermometers made and used by the members. One of these old thermometers was given by the Grand Duke of Tuscany to the late Prof. Babbage, and is now in the Cavendish Laboratory at Cambridge. In England about the same time Boyle made experiments on thermometers. His "Lectures on Cold" were published in 1665 in obedience to the command of the Royal Society, "imposed on me in such a way that I thought it would less misbecome me to obey it unskilfully than not at all. Especially since from so illustrious a company (where I have the happiness not to be hated) I may, in my endeavours to obey and serve them, hope to find my failings both pardoned and made occasions of discovering the truth I aimed at."

From *Nature* 9 May 1901.

50 YEARS AGO

The publication of the Sanskrit manuscript of "Mānasōllāsa" by King Sōmēśvara, son of King Vikramāditya VI of the later Chālukyas, in the Vikramāditya VI of the later Chālukyas, in the Gaekwad's Oriental Series (Publication No. XXVIII, Baroda, 1925), has brought to light a chapter on angling which gives details, almost modern in practice, about this pleasant pastime. "Mānasōllāsa" is an encyclopædic work and was composed in A.D. 1127. The kingdom of Sōmēśvara, comprised practically the whole of the Deccan plateau and included the Godavari, Nerbada, Tapi and Kistna river systems.... The work referred to here shows that the art of angling was developed in ancient India to a very high standard, for the methods described therein are quite in line with those used by anglers in India to-day. The gipsies of Europe, who use Mongolian, Hindi and other fragments of Asiatic languages, to-day practise the same methods as those described by King Sōmēśvara, and it is likely that they wandered from India to Europe and spread the art of angling there.

From *Nature* 12 May 1951.

tions in methylation and chromatin structure, illustrating how changes in epigenetic regulation can spiral out of control towards genomic instability. ■

David E. Symer is in the Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, and the Division of Medical Oncology, Johns Hopkins Oncology Center, 725 North Wolfe Street, Baltimore, Maryland 21205, USA.

Judith Bender is in the Department of Biochemistry and Molecular Biology, Johns Hopkins University School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA.
e-mail: dsymer@jhmi.edu

Evolutionary genomics

Sex and the X

Laurence D. Hurst

Are genomes made up of strings of genes in no particular order? It seems not, given the abundance on the mouse sex chromosomes of genes involved in the manufacture of sperm.

It is often presumed that there is no relationship between what a gene does and where it is in the genome. Nonetheless, there are more genes involved in the manufacture of sperm (spermatogenesis) on the mammalian Y chromosome than might be expected by chance¹. Now, writing in *Nature Genetics*, Wang *et al.*² describe how they identified 25 genes involved in spermatogenesis in mice and found that they are over-represented on both the X and Y chromosomes. This finding adds to the growing realization that genes on the X and Y chromosomes are commonly involved in the development of differences between the sexes.

This might at first sight appear self-evident. Indeed, the mammalian gene that makes males male (*Sry*) must be Y-linked as its presence defines the Y chromosome. Likewise, genes might be on the Y chromosome simply as a convenient way of ensuring that they are expressed only in males. But it is not at all obvious why a gene that is expressed only in males need be on the X chromosome. Something must act to turn the gene on only in males, whether on the X chromosome or on an autosomal (non-sex) chromosome. Given the presence of such a switch mechanism, what is gained by having the gene on the X rather than on an autosome? That the X chromosome might contain an excess of genes that are sex specific in their expression was predicted almost 20 years ago³, however.

To isolate the genes involved in spermatogenesis, Wang *et al.*² used a subtraction method. The cells they looked at were spermatogonia, which are found in the testis and are involved in the early stages of sperm production. These cells have a full complement of genes, as opposed to the half-

1. Miura, A. *et al.* *Nature* **411**, 212–214 (2001).
2. Singer, T., Yordan, C. & Martienssen, R. *Genes Dev.* **15**, 591–602 (2001).
3. The Arabidopsis Genome Initiative *Nature* **408**, 796–815 (2000).
4. Yoder, J. A., Walsh, C. P. & Bestor, T. H. *Trends Genet.* **13**, 335–340 (1997).
5. Federoff, N. *Cell* **56**, 181–191 (1989).
6. Martienssen, R. & Richards, E. *Curr. Opin. Genet. Dev.* **5**, 234–242 (1995).
7. Bird, A. *Trends Genet.* **13**, 469–470 (1997).
8. Jeddeloh, J., Stokes, T. & Richards, E. *Nature Genet.* **22**, 94–97 (1999).
9. Hirochika, H., Okamoto, H. & Kakutani, T. *Plant Cell* **12**, 357–368 (2000).
10. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. *Nature Genet.* **20**, 116–117 (1998).
11. Jackson-Grusby, L. *et al.* *Nature Genet.* **27**, 31–39 (2001).
12. Whitelaw, E. & Martin, D. I. K. *Nature Genet.* **27**, 361–365 (2001).

complement found in sperm themselves. In these experiments a sample of gene transcripts from spermatogonia was depleted of transcripts shared with non-germline cell types. Wang *et al.* then determined the DNA sequence of the residual, spermatogonia-specific transcripts. They identified a total of 36 genes, including all of the eight previously known to be expressed exclusively in spermatogonia. Of these 36 genes, 25 were confirmed to be expressed in spermatogonia only; the remaining 11 made for an informative comparison sample of genes, as they were expressed in both male and female germ cells.

Where are these genes in the genome? The two sets told different stories. Of the 11 that are expressed in both sexes, one is X-linked and ten are on the autosomal chromosomes. Nothing unusual there. By contrast, 10 of the 25 spermatogonia-specific genes are X-linked and three are Y-linked (Fig. 1). If the 22 non-Y-chromosome, spermatogonia-specific genes were randomly distributed, then we would expect only about one to be X-linked.

This finding has at least one ramification. Sometimes, in both mammals and fruitflies, matings between different species result in normal daughters but sterile sons. Many of the hybrid sterility genes are X-linked⁴ but quite why is a matter of contention. That the X chromosome contains so many genes involved in the manufacture of sperm must now be considered as a likely component of the explanation.

More generally, the finding accords with other studies that suggest that X-linked genes are commonly involved in determining the differences between the sexes. For example, analysis of human disease mutants reveals that a gene locus on the X chromosome is at least three times more likely to be involved in sexual development than is a locus on an autosomal chromosome⁵. Similarly, the average influence of X-linked genes on features that are used by males to attract females is much larger than the average influence of X-linked genes on other traits⁶.

This general result, that sex chromosomes harbour many sex-related genes, was predicted by Rice³. He considered the fate of 'sexually antagonistic' alleles — that is, different forms of the same gene that are good for one sex but bad for the other. An example could be an allele that makes individuals brightly coloured. This may benefit males by

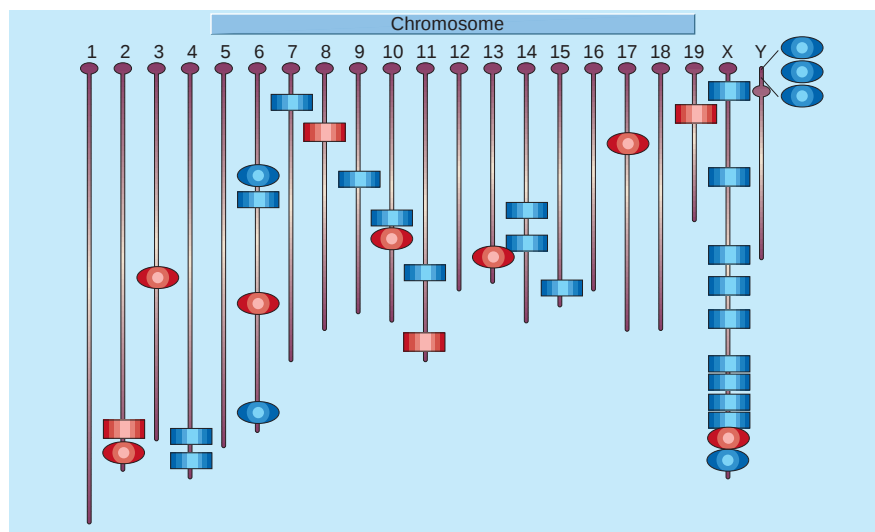


Figure 1 Location in the mouse genome of 36 genes active during sperm production. Genes in blue are expressed in the spermatogonia (male germ cells) only; those in red are expressed in the ovary (female germ cells) as well. Wang *et al.*² find that the spermatogonia-specific genes are heavily concentrated on the X chromosome (10 out of 25). Previously known genes are shown as ovals, and those newly identified by Wang *et al.* as oblongs. (Modified from ref. 2.)

making them conspicuous (and attractive) to females, but is disadvantageous to females as it simply increases the chance that predators will detect them.

Consider what will happen to an initially rare recessive allele that is beneficial to males (the XY sex) but deleterious to females (the XX sex). If it is located on the X chromosome, the allele is very likely to increase in frequency, even if the benefit to males is considerably less than the cost to females. This is because in the XY condition, the beneficial effects are evident, whereas in females the detrimental effects are masked when the allele is rare. By contrast, the allele is unlikely to increase in frequency if it is on an autosomal chromosome, for two reasons. First, because it is recessive, its effects would be hidden. Second, even if they are not perfectly hidden, the net advantage to the males must outweigh the net detriment to the females. The X chromosome therefore acts, overall, as a filter for sexually antagonistic alleles.

Furthermore, as the allele on the X chromosome rises in frequency in the population, some females will carry two copies of it. In this condition the effects of the allele will not be masked and will be detrimental to females. Natural selection should then favour abolition of its expression in females³. Rice's model therefore makes good sense of the fact that genes expressed exclusively in male germ line are predominantly X linked, but those expressed in both sexes are not.

The model can also make sense of the accumulation^{1,2} of spermatogenesis genes on the Y chromosome. In this case, the argument is that, because Y-linked genes occur exclusively in males, negative effects in females need never be realized. Such a mechanism could also explain⁷ why, in guppies, many of the genes for male-specific traits (such as coloration) that females find attractive are on the Y chromosome⁸.

Numerous uncertainties remain, however. Rice's model proposes that the spermatogenesis genes on the X chromosome evolved their function and male-specific expression after becoming X-linked. But could it be the other way around: could genes be more likely to become X-linked, possibly by chromosomal translocation, because they are ancestrally sexually antagonistic⁹? Analysis of the expression of the same genes in non-mammalian vertebrates could resolve the issue. Further, muscle-specific genes¹⁰ are especially abundant on the X chromosome, whereas placentally expressed ones¹¹ are unusually rare. It is not clear why¹².

These uncertainties aside, it is becoming clear that genomes are not the randomly arranged set of genes that we might have imagined them to be. But is this also true at a smaller scale? For example, could natural selection explain which genes are next to each other on a given chromosome? The clustering of broadly expressed genes¹³,

such as those for the everyday running of cells, suggests that small-scale gene arrangements are not always random. These could be exceptions rather than the rule, however, and clarification of the issues awaits the sequencing of the complete genomes of more organisms. ■

Laurence D. Hurst is in the Department of Biology and Biochemistry, University of Bath, Bath BA2 7AY, UK.

e-mail: l.d.hurst@bath.ac.uk

1. Lahn, B. T. & Page, D. C. *Science* **278**, 675–680 (1997).
2. Wang, P. J., McCarrey, J. R., Yang, F. & Page, D. C. *Nature Genet.* **27**, 422–426 (2001).

3. Rice, W. R. *Evolution* **38**, 735–742 (1984).
4. Coyne, J. A. & Orr, H. A. in *Speciation and its Consequences* (eds Otte, D. & Endler, J.) 180–207 (Sinauer, Sunderland, MA, 1989).
5. Saifi, G. M. & Chandra, H. S. *Proc. R. Soc. Lond. B* **266**, 203–209 (1999).
6. Reinhold, K. *Behav. Ecol. Sociobiol.* **44**, 1–7 (1998).
7. Fisher, R. A. *Biol. Rev.* **6**, 345–368 (1931).
8. Winge, O. *J. Genet.* **18**, 1–43 (1927).
9. Charlesworth, D. & Charlesworth, B. *Genet. Res.* **35**, 205–214 (1980).
10. Bortoluzzi, S. *et al. Genome Res.* **8**, 817–825 (1998).
11. Ko, M. S. H. *et al. Hum. Mol. Genet.* **7**, 1967–1978 (1998).
12. Hurst, L. D. & Randerson, J. P. *Trends Genet.* **15**, 383–385 (1999).
13. Caron, H. *et al. Science* **291**, 1289–1292 (2001).

Earthquakes

Shock delay

Elizabeth Harding Hearn

Postseismic stress transfer is the local redistribution of stresses in the Earth's crust that follows an earthquake. It's a hot topic at the moment, and understandably so because of its relevance in assessing seismic hazards.

The magnitude-7.1 Hector Mine earthquake occurred only about 20 km from the site of another large earthquake in southern California, but happened seven years later. Were the two connected? On page 180 of this issue¹, Freed and Lin describe their investigations into the question.

Earthquakes change stresses in the surrounding Earth's crust, so it is not surprising that the initial jolt of a major earthquake is followed by a sequence of aftershocks that (fortunately) decay with time. Aftershocks tend to occur on faults where the 'shear stress' (τ) loading the fault has increased or the 'compressive stress' (σ) clamping the fault has decreased, or both². The Hector Mine earthquake of 1999 did not seem to fit into this picture: it occurred geographically close to, but seven years after, the magnitude 7.3 Landers earthquake, and in an area where most researchers consider that the change in stress after Landers should not have encouraged seismic activity.

Freed and Lin¹ now show how postseismic deformation — continued deformation of the Earth's crust during the years after the Landers earthquake — may have increased stresses at the hypocentre of the Hector Mine earthquake and triggered that event. Their modelling study shows that, within a few years of an earthquake, postseismic deformation may change stresses in the upper crust by at least as much as the stress released by the earthquake itself. The study shows, too, that estimating such changes is crucial for characterizing regional seismic hazards.

Seismologists have been fairly successful at explaining the spatial² and temporal³ distribution of aftershocks using models that account only for the stress changes occurring

during an earthquake (coseismic deformation). According to these models, an interplay between the Coulomb stress (τ minus a scaling constant times σ) and the time-dependent evolution of frictional properties of faults explains how an earthquake can trigger other earthquakes within about a hundred kilometres for days to months after the main shock. But sustained postseismic deformation is required to explain the triggering of aftershocks that occur several years after a large earthquake. If the entire crust were effectively elastic (that is, it did not undergo distributed deformation over time following a coseismic stress change), any broadly distributed postseismic deformation would probably be due to sliding of faults at depth^{4,5}. But the lower crust and upper mantle are viscoelastic materials that cannot support significant shear stresses and tend to yield with time. After an earthquake, the lower crust gradually relaxes its grip on the upper crust, allowing the upper crust to continue deforming in a directional sense comparable (in most areas) to the coseismic deformation.

Freed and Lin¹ propose that viscoelastic relaxation of the lower crust⁶ continuously increased the Coulomb stress at the Hector Mine hypocentre, until it exceeded a threshold and triggered the earthquake. They find that the postseismic Coulomb stress change at the hypocentre was about 1 bar, which is several times the most generous published estimate of coseismic Coulomb stress⁷. Viscoelastic relaxation efficiently transmits postseismic stresses to great distances, so Freed and Lin's model also suggests that the Landers earthquake has resulted in considerable changes in stress on other active faults in southern California, such as the San Andreas.

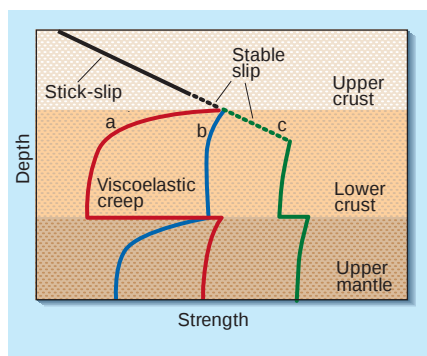


Figure 1 Models of postseismic deformation: strength versus depth profiles for models differing in assumed crust and mantle viscosities. a, Low-viscosity lower crust^{1,6}; b, low-viscosity upper mantle⁸; c, high-viscosity lower crust and upper mantle, with early postseismic deformation accommodated by frictional slip on a fault⁴ or creep in a relatively weak shear zone⁹. 'Strength' is either viscosity, or frictional resistance to breaking or sliding (which may vary during the earthquake cycle). Viscoelastic deformation may occur throughout the crust or upper mantle (models a and b) or, if the crust and mantle are too strong to deform significantly after the earthquake, it may be localized along a shear zone with low viscosity (model c). Alternatively, postseismic deformation along a surface (model c) may be due not to viscoelastic creep (which requires high temperatures) but to frictionally controlled slip on a fault that does not break coseismically ('stable slip'). Where a fault fails during an earthquake, it exhibits 'stick-slip' frictional behaviour. The evolution of upper crustal stresses after an earthquake (and of seismic hazard associated with nearby faults) depends on whether a, b, c or another model applies.

Freed and Lin also test an alternative model in which post-Landers deformation is attributed to relaxation of the mantle⁸, rather than the lower crust. They conclude that the Coulomb stress at the Hector Mine hypocentre increases significantly regardless of the depth to the relaxing viscoelastic layer. Such an exploration of model sensitivity is crucial given the uncertainty about how the viscosities of continental crust and upper mantle vary with depth. Another feature of their paper is its exposition of how sensitive coseismic stresses near a fault may be to the distribution of coseismic slip. Coseismic stresses provide the driving force for subsequent postseismic deformation, so model errors in the coseismic stress pattern will continue to manifest themselves over time. This phenomenon illustrates the importance of evaluating how differences between various elastic models have arisen, and of choosing a favoured one, as Freed and Lin have done.

Uncertainty over viscosities in the lower crust and upper mantle leads to another controversy — that of how much of the

postseismic stress change in the crust is due to delayed slip on fault surfaces (frictionally controlled afterslip), time-dependent deformation concentrated along shear zones (viscoelastic creep) or other processes. If the viscosities of the lower crust and upper mantle are high, viscoelastic deformation will occur so slowly that another process, such as frictional afterslip, must be responsible for most postseismic deformation during the months to years after an earthquake (Fig. 1). Also at issue are whether the viscoelastic deformation rate is proportional to stress, or to stress to some power greater than 1 (as indicated by many rock-deformation experiments), and how low-viscosity material is distributed in the crust. Earth scientists must resolve these questions because various models predict different patterns of change in postseismic stress, and thus in seismic risk, even when the models yield similar patterns of horizontal surface deformation near the earthquake rupture.

To distinguish between the different models we will need more data on postseismic deformation, both in time and in space. Monitoring programmes are under

way in many parts of the world, and proposed initiatives such as Earth Scope (currently being considered by the US National Science Foundation) promise to provide more detailed measurements of three-dimensional postseismic surface displacements and strain. Along with information from seismic studies, geological mapping and laboratory experiments, modelling work using these measurements will help to reveal more about earthquake cycles, and allow for near-real-time estimates of the build up and release of stress on active faults.

Elizabeth Harding Hearn is in the Department of Earth, Atmospheric and Planetary Sciences, 54-614, Massachusetts Institute of Technology, Boston, Massachusetts 02139-4307, USA.

e-mail: lizh@chandler.mit.edu

1. Freed, A. M. & Lin, J. *Nature* **411**, 180–183 (2001).
2. King, G. C. P., Stein, R. S. & Lin, J. *Bull. Seismol. Soc. Am.* **84**, 935–953 (1994).
3. Dieterich, J. *J. Geophys. Res.* **99**, 2601–2618 (1994).
4. Tse, S. T. & Rice, J. R. *J. Geophys. Res.* **91**, 9452–9472 (1986).
5. Dieterich, J. H. *J. Geophys. Res.* **84**, 2161–2168 (1979).
6. Deng, J. *et al. Science* **282**, 1689–1692 (1998).
7. Parsons, T. & Dreyer, D. *Geophys. Res. Lett.* **27**, 1949–1952 (2001).
8. Pollitz, F. F., Peltzer, G. & Bürgmann, R. *J. Geophys. Res.* **105**, 8035–8054 (2000).
9. Sibson, R. H. *J. Geol. Soc. Lond.* **140**, 741–767 (1983).

Computational biology

Beyond the spherical cow

John Doyle

Computational and mathematical models are helping biologists to understand the beating of a heart, the molecular dances underlying the cell-division cycle and cell movement, and much more.

The sequencing of the human genome has inspired many to anticipate a time when the most complicated biological phenomena will be understood in detail. But sceptics worry that a simple list of genes and proteins, no matter how complete, will be as useless to biologists as a list of a computer's transistors would be to an understanding of the software used to browse the online version of this article. At a recent meeting*, however, both sides found cause for optimism in computational and mathematical approaches to complement experiments in unravelling complex biological processes.

Computation is so prominent in genomics that the naive observer could be excused for mistaking this field for a branch of robotics or computer science. But the computational approaches discussed at the meeting were firmly focused on post-genomic issues — the dynamics and control of the networks of genes and proteins at work in cells. Here, increasingly sophisticated software packages, such as those at the web links listed at the end

of this article, are proving vital. With them, biologists are finally gaining access to technology that has hitherto been restricted to the design of complex engineering systems, such as integrated circuits, aircraft, communication networks and chemical-processing plants. Many experimentalists are using mathematical modelling and computation to study biological systems that are just too complex to be understood by intuition and informal models. Meanwhile, those who are developing computational models fully appreciate that their models must be integrated with experimental data.

The control of the cell-division cycle is one area where complex molecular mechanisms, worked out by legions of molecular biologists, have been codified into a single mathematical model (J. Tyson, Virginia Polytechnic Inst.). Building such a model requires a long process of data collection and verification, but the efforts are well worth it. For example, models like this can be used to test theories about cellular processes by comparing the effects of virtual and real mutated genes.

The most popular mathematical models, in this and other studies, involve ordinary and partial differential equations to rep-

*First International Symposium on Computational Cell Biology. Cranwell Resort, Lenox, Massachusetts, USA. 4–6 March 2001. www.nrcam.uchc.edu/conference

resent chemical kinetics. But models that systematically integrate several scales are also important. For example, stochastic chemical kinetics (D. Bray, Univ. Cambridge) can be used to model individual molecular interactions at various levels of detail. Such models capture fluctuations that tend to clump together in models based simply on differential equations. Also, the chemical dynamics of cells must be integrated with mechanical aspects (D. Ingber, Harvard Med. School), which — just as in complex engineering systems — become increasingly important at the network level. In another parallel to engineering principles, a model that integrates hypersensitivity and positive feedback in a signalling pathway known as the MAP kinase cascade predicts the occurrence of switch-like behaviour, where some internal chemical state of the pathway switches, as in a digital system (J. Ferrell, Stanford Univ.). Such behaviour is uncovered experimentally only when single cells are analysed (which is not common).

Widespread interest in sophisticated software is new. But the general approach of integrated modelling has a long and successful history. Computer modelling of individual ion channels and transporters in cardiac cells, at the level of differential equations, goes back 40 years. Since then, models of heart function have incorporated further details, such as pacemaker activity, the genetic defects underlying arrhythmic heart beats, mechanical–electrical feedback, and regional patterns of expression of ion transporters, to the point where high-fidelity simulations of the whole heart are possible (D. Noble, Oxford Univ.). Iterative interactions between experiment and simulation, and between molecular mechanisms and the physiology of the whole organ, were essential to this success.

Speaking from the experimentalist's point of view, T. Pollard (Salk Inst., San Diego) described how computation fits into a broad agenda for analysing cell-biological data. He detailed a model for cell motility, which involves actin filaments — a type of cytoskeletal structure. The model was constructed from a huge amount of experimental knowledge, including quantitative data on the identities, affinity constants and reaction rates of the interacting molecules. Such data are collected only rarely, but will become increasingly important.

Several other areas of biology have benefited from the interaction of experiment and modelling. For example, reaction–diffusion models (standard models of spatially dependent chemical kinetics) formed the basis for the first detailed picture of the transport of molecules into and out of the nucleus (A. Smith and I. Macara, Univ. Virginia). Likewise, the transport and sorting of proteins between the endoplasmic reticulum and the Golgi complex have been analysed by com-

binning modelling with studies involving fluorescent markers (J. Lippincott-Schwartz, NIH, Bethesda). The transport of fluorescently labelled RNA granules to cellular sites of protein synthesis can be broken down into three characteristic patterns — vibrations, oscillations and movement. A stochastic model, integrating the rates at which motor proteins bind to these granules and are activated, can account for all of these behaviours (J. Carson, Univ. Connecticut Health Center).

Finally, the availability of electrophysiological and fluorescence methods to obtain quantitative data has made the dynamics of calcium ions one of the most popular subjects for modelling. A new model of calcium 'sparks' — the release of calcium ions from an organelle known as the sarcoplasmic reticulum — can account for the previously mysterious 'termination' phase of a spark (J. Lederer, Univ. Maryland). This model involves interactions within a cluster of calcium channels, and regulation of the channels by calcium within the sarcoplasmic reticulum. Sparks in skeletal muscle can be studied in preparations from frog muscle, and have been further analysed with a model that incorporates the characteristics of the microscope used to gather the data (D. Uttenweiler, Univ. Heidelberg). An image-based model of calcium waves in a neuron (L. Loew, Univ. Connecticut Health Center) emphasized the role of cell geometry, which can control the spatial and temporal patterns of calcium ions.

A prominent theme was that biology needs more theory, in addition to modelling and computation, to make sense of complex networks. Without underlying theories, software would be of limited use in engineering systems. Theory has a rather bad reputation among biologists, in part because the ideas so popular in physics — such as pattern formation, critical phase transitions and chaos — have proved largely irrelevant to molecular biology. So far, engineering theories of communication, control and computing have had little contact with biology, but perhaps that should change. If biologists are much like physicists in stretching the limits of experimental reductionism, they are also like engineers in revelling in the enormity, variety and sheer complexity of the systems they study. No interest in spherical cows here. ■

*John Doyle is in the Department of Control and Dynamical Systems, California Institute of Technology, Pasadena, California 91125, USA.
e-mail: doyle@cds.caltech.edu*

Web links

- ♦ <http://www.cds.caltech.edu/erato>
- ♦ <http://websites.ntli.com/~igor.goryanin>
- ♦ <http://www.e-cell.org>
- ♦ <http://www.gepasi.org>
- ♦ <http://members.tripod.co.uk/sauro/biotech.htm>
- ♦ <http://www.zoo.cam.ac.uk/comp-cell/stochsim.html>
- ♦ <http://www.nrcam.uchc.edu>

Daedalus

The art of spinning

The gecko, that engaging lizard that climbs walls and runs across ceilings, is alleged to do so by means of the 'setae' on its feet — tiny hairs which stick to its anchorage by van der Waals forces. Daedalus does not believe it. A crumbly wall or powdery ceiling would easily defeat such forces, which occur only between surfaces in very intimate contact.

Similarly, engineering slip gauges, which are ground to better than 0.01 millimetre, cling together when 'wrung' into a pile for measurement. Again, Daedalus absolves van der Waals forces, and atmospheric pressure, from this feat. He reckons that the thin film of oil on the gauges is responsible, hauling them together by capillary forces. If the gauges were carefully dried by means of dichloromethane, for example, and then heated to about 35 °C (the boiling point of that solvent), they should fall apart.

In fact, Daedalus reckons that the only application of van der Waals forces in ordinary life is the ancient art of spinning. The business of taking short fibres, and twisting them together into a long, strong thread, is most remarkable. The fat on wool, and the oil on cotton, must play an initial role, tying the strands together. Yet after a while van der Waals forces must play their part, for the fabric does not fall apart again on washing.

So Daedalus proposes a new spinning technology. An initial thread, maybe a wire or carbon fibre, or even a traditional yarn, will be raised to a high voltage while other porous materials are released around it. Expanded polystyrene, paper pieces, cotton, wool and so on will be released nearby and will stick firmly by coulombic forces, which must also bring the surfaces into intimate contact. The resulting 'open fibre' should be as strong as anything van der Waals forces could bring about. Daedalus wonders whether oil will be needed to reinforce the initial bond before the fibre is converted to a fabric and washed.

All sorts of materials can be floated onto open fibre: fumed silica, fine metal wires, even expanded vermiculite. Daedalus has set his sights on a fibre as strong as conventional ones, but with far more 'bulk' and thermal resistance. It will knit or loom-weave into an amazingly warm cloth. DREADCO engineers have free rein to add any bulky insulator to the fibre, in the hope of improving these properties. Daedalus himself hopes to gain a bit of insight into how ordinary spinning actually works.

David Jones

Spiny lobsters stick and slip to make sound

These crustaceans can scare off predators even when their usual armour turns soft.

Many arthropods are able to produce pulsed sounds by rubbing a hard pick over stiff macroscopic ridges¹, rather like dragging a stick over a washboard. Spiny lobsters (Palinuridae) also make pulsed sounds, and here I show that they generate these by virtue of a frictional 'stick-and-slip' mechanism that is more usually associated with bowed stringed instruments. By using this technique rather than a 'hard-washboard' mechanism, lobsters can produce strident warning sounds against predators throughout their moult cycle, including the period when their exoskeleton is softened and they are most susceptible to predation.

Palinurid lobsters produce a loud, abrasive sound (rasp) by rubbing a plectrum (a basal extension of each antenna) over a file (located on the antennular plate below the eyes; Fig. 1)². This rasp, which is composed of a series of sound pulses, is ostensibly similar to stridulation-based sounds made by crickets, for example, in which a sound pulse is produced when a hard pick (plectrum) hits a macroscopic ridge on the file (Fig. 2a)¹. However, lobsters produce sound pulses without impact between hard structures; instead, a soft-tissue plectrum rubs over microscopic shingles on the file (Fig. 2b).

The lobster plectrum and file suggest a frictional mechanism³ that is analogous to the 'stick-and-slip' that occurs between the bow and string of stringed instruments⁴. In stringed instruments, friction causes the bow to stick momentarily and then to slip relative to the string many times during the course of a single sweep of the bow across the string. This unsteady movement of the bow relative to the string excites vibrations and produces sound. Although theoretically possible in biological systems⁵, such a mechanism has not been found until now.

In the lobster stick-and-slip model,



Figure 1 'Face' view of *Panulirus ornatus*, which produces sound using a stick-and-slip mechanism. Eyes are either side of the central orange panel; files are directly below eyes and plectra are the pink structures underneath. Dorsal is towards the top of the page; posterior is into the page.

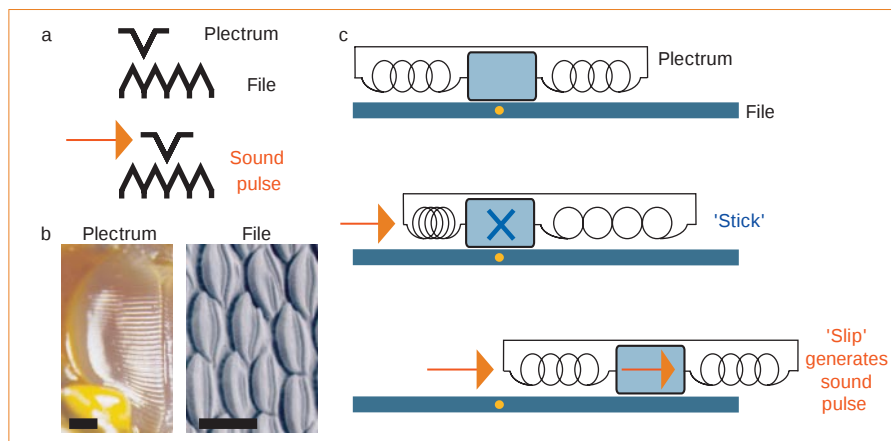


Figure 2 Comparison of the 'stick-and-slip' mechanism used by a spiny lobster (*Panulirus argus*) to produce sound and the 'washboard' mechanism typically used by other arthropods. **a**, In typical stridulators such as the cricket, each sound pulse is produced when a hard pick on the plectrum hits a ridge on the file. **b**, Morphology of the sound-production surfaces of the spiny lobster. Soft-tissue plectrum ridges (left) are rubbed posteriorly (to the right as shown) over the anteriorly projecting shingles on the file (right). Scale bars, 500 μ m (left) and 15 μ m (right). **c**, In the stick-and-slip model, a series of sound pulses is generated when the static friction between the plectrum and file surfaces is periodically exceeded by the sliding friction of the plectrum as it is pulled across the file. Detailed methods are available from the author.

sliding friction between the plectrum and file surfaces would intermittently exceed static friction as the two rub together (Fig. 2c). The soft, elastic tissue of the plectrum resists compression and hence would store energy during the 'sticking' phase and release it during the 'slipping' phase. Each slip between the two surfaces would thus generate a sound pulse.

To test whether palinurid lobsters produce sound using such a stick-and-slip mechanism, I first measured the correlation between plectrum movement and sound production. Using the Caribbean spiny lobster *Panulirus argus*, I attached a motion detector to the plectrum and used a hydrophone to record rasps. Sound was generated only when the plectrum moved posteriorly against the anteriorly projecting shingles on the file. This posterior movement consisted of alternating still and sliding periods; sound pulses were produced only during the sliding movements (97 rasps; 6 lobsters). Using high-speed video analysis, I found that the frequency of shingle impacts was 19,000–28,000 Hz (5 rasps; 1 lobster), which far exceeds the average pulse rate (77 Hz; 139 rasps; 6 lobsters). This is inconsistent with a 'pick-and-washboard' mechanism, in which the pulse rate would be equal to the rate of ridge impact.

A single continuous muscle contraction should be sufficient to produce the series of stepped movements described here, just as a constant driven motion of a bow over a violin string produces repeated sticking and

slipping⁴. Synchronous electromyographic, acoustic and plectrum-movement recordings showed that the promotor muscle⁶ contracts tonically during the rasp (120 rasps; 6 lobsters). Thus, a single tonic muscle contraction generates a series of sound pulses. Furthermore, pulsed sound can be generated by manually dragging the plectrum over the file with a single pull. These results are further evidence for the presence of a stick-and-slip mechanism.

When attacked by predators, spiny lobsters produce rasps^{3,7,8}, which probably function as a startling deterrent⁹. Lobsters typically rely on their hardened exoskeletons for protection from attack, but they lack such protection when this is softened during the moult cycle¹⁰. Lobsters do not depend on the hard washboard structures typically used by arthropods to produce sound, but instead make use of frictional interactions between surfaces that do not need to be hard. They can therefore create loud sounds even with a softened exoskeleton during the moult cycle¹¹, deterring predators acoustically when their other structural defences are compromised.

Sheila N. Patek

Department of Biology, Duke University, Durham, North Carolina 27708, USA

e-mail: snp2@duke.edu

1. Ewing, A. W. *Arthropod Bioacoustics: Neurobiology and Behaviour* (Cornell Univ. Press, Ithaca, 1989).
2. Phillips, B. F., Cobb, J. S. & George, R. W. in *The Biology and Management of Lobsters: Physiology and Behavior* (eds Cobb, J. S. & Phillips, B. F.) 1–82 (Academic, New York, 1980).
3. Meyer-Rochow, V. B. & Penrose, J. D. *J. Exp. Marine Biol. Ecol.* **23**, 191–209 (1976).

4. Benade, A. H. *Fundamentals of Musical Acoustics* (Dover, New York, 1990).
5. Fletcher, N. H. *Acoustic Systems in Biology* (Oxford Univ. Press, New York, 1992).
6. Paterson, N. F. *Ann. S. Afr. Mus.* **51**, 1–232 (1968).
7. Mulligan, B. E. & Fischer, R. B. *Crustaceana* **32**, 185–199 (1977).

8. Lindberg, R. G. *Univ. California Publ. Zool.* **59**, 157–248 (1955).
9. Masters, W. M. *Behav. Ecol. Sociobiol.* **5**, 187–200 (1979).
10. Aiken, D. E. in *The Biology and Management of Lobsters: Physiology and Behavior* (eds Cobb, J. S. & Phillips, B. F.) 91–163 (Academic, New York, 1980).
11. Moulton, J. *Biol. Bull.* **113**, 286–295 (1957).

Plant development

Signals from mature to new leaves

Stomata are microscopic pores on the surfaces of leaves, the number and density of which vary in response to changes in environmental conditions, such as carbon dioxide concentration and light. We show here that mature leaves of *Arabidopsis thaliana* detect and transmit this external information to new leaves of the same plant, producing an appropriate adjustment of stomatal development. As CO₂ concentration controls both stomatal opening¹ and number^{2,3}, and stomatal numbers also increase with higher light intensity⁴, the large gradients of CO₂ and light found within plant communities⁵ have the potential to influence stomatal development.

Every year, 40% of the CO₂ in the atmosphere passes through stomata⁶. Stomatal numbers modify both photosynthesis and efficient use of water¹, and so any change in stomatal numbers⁷ in response to CO₂ and light can influence photosynthesis and atmospheric CO₂ concentration. The response mechanism described here may exert global effects that are not currently

included in canopy and vegetation models of increasing atmospheric CO₂ levels.

A genetic component in the response of stomatal development to increasing atmospheric CO₂ concentration has been identified⁸ and the control of stomatal development by CO₂ concentration is known to occur during early leaf development^{9,10}, when ambient CO₂ concentrations may not be accurately detectable by a new leaf sheathed by antecedent leaf primordia¹¹. We therefore tested whether CO₂ concentration can be detected by mature leaves in open ambient conditions, which might then transmit a signal to induce an appropriate developmental response by the stomata of new leaves. Our experimental design using the model plant *Arabidopsis thaliana* is shown in Fig. 1a.

Expanding leaves outside the cuvette (ambient CO₂, 360 p.p.m.), with mature leaves exposed to a high concentration of CO₂ inside the cuvette (720 p.p.m.; Fig. 1a), developed with a reduced stomatal index and density (Fig. 1b, left) compared with control plants grown entirely at ambient CO₂. Reversing the cuvette arrangement so that the mature leaves were exposed to 360 p.p.m. and the expanding leaves to 720 p.p.m. CO₂ resulted in a complete reversal

(increase) of the stomatal index and density of the new leaves (Fig. 1b, right).

In these experiments, both abaxial (upper) and adaxial (lower) leaf surfaces responded in a similar manner, indicating that CO₂ concentration is detected by mature leaves which signal to expanding leaves to induce an appropriate developmental response. To our knowledge, this is the first demonstration that mature leaves both detect CO₂ concentration and transmit a long-distance signal that controls stomatal development in young leaves. Expanding leaves appeared to have no capacity to detect ambient CO₂ concentration or to respond to it directly by altering stomatal initiation; no cuvette effect was found to contribute to the responses.

This mechanism of CO₂ detection and signalling could enhance or optimize performance in plant communities. For example, the reduction in stomatal index and density with CO₂ enrichment enhances water-use efficiency^{3,12}, but such a response will be much less effective¹³ when leaves develop in the shade of other plants. We therefore tested whether the mechanisms controlling the response of stomatal initiation to CO₂ might also be accompanied by another that responds to light, reducing the initiation of stomata with increasing shade.

We have previously shown that there is a positive relation in *Arabidopsis* between irradiance and both stomatal index and density. Fully grown leaves were now placed in shaded light, with the expanding leaves under full light. The stomatal index (Fig. 1c) and density (results not shown) of new leaves were reduced, as if they had developed under shade conditions. We conclude that long-distance signalling must also be involved in controlling the response of stomatal development to light, which supports the idea that there is an ecologically important link between the responses evoked by light and by carbon dioxide.

J. A. Lake, W. P. Quick, D. J. Beerling, F. I. Woodward

Department of Animal & Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK
e-mail: F.I.Woodward@Sheffield.ac.uk

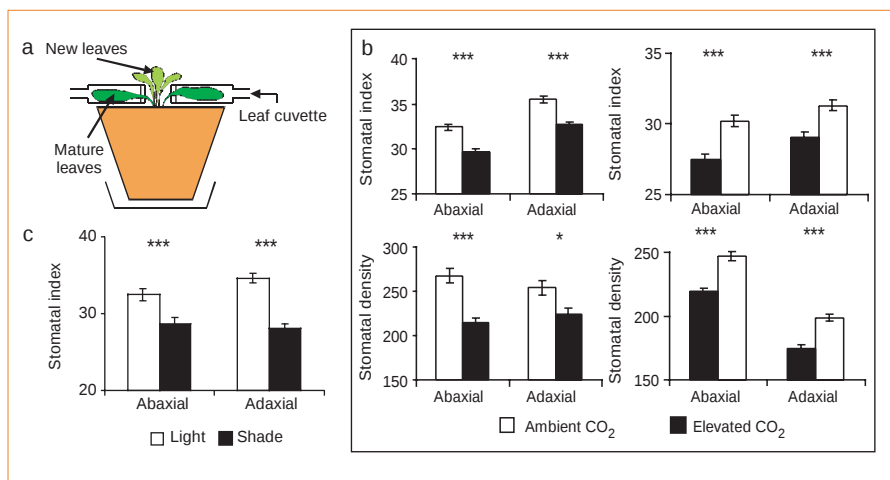


Figure 1 Mature leaves detect changes in CO₂ concentration and elicit a stomatal response in developing leaves. **a**, Leaf-cuvette experiment. Plants of *Arabidopsis* (Columbia, Col-0) were grown for 4 weeks under ambient CO₂ (360 p.p.m.) until leaf insertions 5 to 13 had developed. These mature leaves were enclosed in transparent airtight cuvettes under CO₂ concentrations of either 720 or 360 p.p.m. Subsequent leaf insertions developed outside the cuvette under ambient CO₂. Plants were maintained in cuvettes for 7 to 9 days until the next five leaf insertions had matured, the last three of which were investigated for stomatal density (no. of stomata per mm²) and index ((no. of stomata/no. of stomata + no. of epidermal cells) × 100). **b**, (left) Stomatal index and density for new leaves (insertions 16 to 19) under ambient CO₂ when mature leaves (insertions 5 to 13) inside cuvettes are supplied with increased CO₂ (720 p.p.m.). Both stomatal density and index are reduced in new leaves if the supply of CO₂ is increased to the mature leaves. Right, reverse experiment: mature leaves inside cuvettes are under CO₂ at 360 p.p.m.; external CO₂ is 720 p.p.m. Stomatal density and index increase in response to the decreased CO₂ around the mature leaves. **c**, Effect on stomatal index of new leaves of reducing light incident on mature leaves by using neutral density filters (shade) or transparent filters (full light). Stomatal index of new leaves is reduced when mature leaves are shaded.

****P* < 0.0005; **P* < 0.05; bars, s.e.m.; *n* = 150.

1. Berryman, C. A., Eamus, D. & Duff, G. A. *J. Exp. Bot.* **45**, 539–546 (1994).
2. Woodward, F. I. *Nature* **327**, 617–618 (1987).
3. Beerling, D. J., McElwain, J. C. & Osborne, C. P. *J. Exp. Bot.* **49**, 1603–1607 (1998).
4. Tichá, I. *Photosynthetica* **11**, 375–471 (1982).
5. Buchmann, N., Kao, W.-Y. & Ehleringer, J. R. *Glob. Change Biol.* **2**, 421–432 (1996).
6. Ciais, P. *et al.* *J. Geophys. Res. Atmos.* **102**, 5857–5872 (1997).
7. Douville, H. *et al.* *J. Geophys. Res. Atmos.* **105**, 14841–14861 (2000).
8. Gray, J. E. *et al.* *Nature* **408**, 713–716 (2000).
9. Woodward, F. I. & Bazzaz, F. A. *J. Exp. Bot.* **39**, 1771–1781 (1988).
10. Donnelly, P. M., Bonetta, D., Tsukaya, H., Dengler, R. E. & Dengler, N. G. *Dev. Biol.* **215**, 407–419 (1999).
11. Jackson, D., Veit, B. & Hake, S. *Development* **120**, 405–413 (1994).
12. Eamus, D., Berryman, C. A. & Duff, G. A. *Aus. J. Pl. Physiol.* **20**, 741–755 (1993).
13. Muraoka, H., Tang, Y., Terashima, I., Koizumi, H. & Washitani, I. *Plant Cell Environ.* **23**, 235–250 (2000).

Biogeology

How old are bacteria from the Permian age?

Discovery of bacteria that remain viable in a dormant state for lengthy periods is significant for understanding patterns of microbial diversity and evolution on Earth, as well as for assessing the possibility of life's interplanetary transport by impact processes. The isolation by Vreeland *et al.*¹ of viable 250-million-year-old bacteria is an extraordinary claim, based on meticulous extraction from evaporite deposits of the Delaware Basin. If valid, this discovery expands dramatically the maximum proposed age for microbial survivability. Here we argue that, although the Permian age of these well-documented deposits is not in question, the fluid inclusions and the viable bacterial spores contained in them may represent much more recent features. The age of these microbes must therefore remain uncertain.

Vreeland *et al.* describe commonly accepted primary evaporite textures and structures — fine-scale fluid inclusions and bedded halite, for example — that are suggestive of the original depositional environment. But these observations are not pertinent to the question at hand because bacterial samples were not obtained from halite displaying such primary features. Instead, bacterial spores were extracted from dissolution pipes of “coarse, clear halite with fewer, but larger, fluid inclusions”. The authors claim that these dissolution pipes are contemporaneous with primary halite, because the coarser crystal pipes “are overlaid by undisturbed (presumably primary) halite beds”: however, this observation is not sufficient to establish the age of the fluid inclusions.

The large, clear, single-crystal nature of the halite selected is not typical of primary halite deposition. Such coarse halite is more commonly associated with processes that occur after — sometimes long after — initial deposition. For example, evaporites of the Delaware Basin have large crystal-lined cavities, which almost certainly formed in a quiet, post-depositional subsurface environment². Coarse halite with fluid inclusions may also form by the dissolution and recrystallization of primary halite. Such recrystallization can occur repeatedly in a salt body through interaction with new pulses of fluid, including bacteria-bearing groundwater from above or below.

Almost all bedded salt contains at least some healed fractures, not always readily visible even by optical microscopy, along which fluids have moved³. These moving fluids may produce pipe-like masses that crosscut many beds — features similar to the those described by Vreeland *et al.* —

and such dissolution and reprecipitation may take place much later than the primary deposit. Superposition of undisturbed salt beds is therefore insufficient to show that the bacteria-bearing halite dissolution pipes and their fluid inclusions are contemporaneous with primary depositional features.

Compositions of fluid inclusions from Delaware Basin evaporites also suggest multiple sources and ages, calling into question the supposed age of 250 million years. Petrographic studies and chemical analysis of large (about 1 mm) fluid inclusions in clear ‘recrystallized’ salt (as sampled by Vreeland *et al.*) show these fluids to be complex bitterns, which result from multiple diagenetic processes at unknown times³. This history is evident from the absolute concentrations as well as the ratios of halogen, alkali and alkaline-earth ions^{4,5}. These compositions vary significantly in adjacent inclusions, often separated by less than 1 mm, and are almost always far from ion ratios obtainable by simple evaporation of sea water. Furthermore, isotope studies of such fluid inclusions from the Delaware Basin suggest that mixture with both ancient and modern meteoric waters has occurred⁶.

We conclude that, in the absence of primary growth features in the specific halite crystals studied, the age of those crystals and their fluids must remain in doubt.

Robert M. Hazen*, Edwin Roedder†

*Geophysical Laboratory and NASA Astrobiology Institute, Carnegie Institution of Washington, Washington, DC 20015, USA

e-mail: hazen@gl.ciw.edu

†47 Salt Island Road, Gloucester, Massachusetts 01930, USA

1. Vreeland, R. H., Rosenzweig, W. D. & Powers, D. W. *Nature* **407**, 897–900 (2000).
2. Shlichta, P. J. *Geol. Soc. Am. Sp. Paper* **88**, 597–617 (1968).
3. Roedder, E. *Am. Mineral.* **69**, 413–439 (1984).
4. Roedder, E., d'Angelo, W. M., Dorrzapf, A. F. Jr & Aruscavage, P. J. *Chem. Geol.* **61**, 79–90 (1987).
5. Bein, A., Hovorka, S. D., Fisher, R. S. & Roedder, E. *J. Sedim. Petrol.* **61**, 1–14 (1991).
6. O'Neil, J. R., Johnson, C. M., White, L. D. & Roedder, E. *Appl. Geochem.* **1**, 265–271 (1986).

Vreeland et al. reply — Hazen and Roedder touch on several geological issues raised by our limited description of the Permian Salado Formation. In our study, we sampled coarse halite in dissolution pipes for testing for viable Permian-age bacteria.

A synsedimentary age of these dissolution pipes is shown by undisturbed overlying beds and by the development of pipes downward from surfaces exposed on desiccated salt pans^{1,2}. Desiccation cracks, dish-shaped salt and clay laminae accumulated in salt polygons or saucers, and insoluble residues all indicate subaerial exposure². Microkarst features (smaller, but similar to dissolution pipes) have been described in cores from the Permian San Andres Formation³, and the undisturbed overlying beds were considered to be evidence of syn-

depositional age of microkarst in Permian halite beds in the Palo Duro Basin, Texas^{3,4}.

Dissolution pipes tend to reach a common depth below an exposure surface². In some beds, macropores (10–30 cm across) filled with coarse halite developed at about that same depth. A common water (brine) table controlled macropores and pipes^{2–4}. Coarse halite filling the synsedimentary dissolution pipes² and microkarst⁴ shows crystal boundary relationships consistent with passive pore-filling cement growth. Some cloudy halite from fine fluid inclusions was found in Salado dissolution pipes¹.

We see no brine conduits through the Salado and know no means of maintaining less than halite saturation in such a case. Permeability decreases quickly in halite beds as halite cements occlude porosity with near-surface crystallization^{4,5}. Based on *in situ* experiments in Salado halite, the undisturbed permeability and hydraulic conductivity are about 10^{-22} m² and 10^{-15} m s⁻¹, respectively⁶. With a hydraulic gradient of 0.01 and a porosity of 0.01, brine would take more than 30 million years to flow one metre. Brine chemistry commonly varies over centimetres, consistent with extremely limited permeability⁷. These characteristics weigh heavily against water flow through the Salado to dissolve and recrystallize halite and against post-Permian natural introduction of bacteria to the halite in the pipes.

We do not assume that sea water is trapped in these fluid inclusions. For example, marine and non-marine Salado beds can be distinguished⁸. Similar bromine concentrations in microkarst and primary halite in the Palo Duro study suggest penecontemporaneous formation from the same brine pool⁵. But differing compositions may not indicate greatly different ages. Variable exposure periods, marine inflows and continental fluid sources lead to complex chemistry, and stable isotopes may resemble mixes of meteoric and evaporated waters⁹, unless Permian rain and sea water were very different from now. Homogeneous fluid and mineral compositions may be better evidence for massive fluid movement through the evaporites.

Salado dissolution pipes are consistent with a synsedimentary origin and Permian age. Synsedimentary dissolution pipes should be useful in diagnosing exposure surfaces within evaporites. We anticipate renewed interest in these and similar deposits as a result of our study.

Dennis W. Powers*, Russell H. Vreeland†, William D. Rosenzweig†

*140 Hemley Road, Anthony, Texas 79821, USA
e-mail: dwpowers@htg.net

†Department of Biology, West Chester University, West Chester, Pennsylvania 19383, USA

1. Powers, D. W. & Hassinger, B. W. *J. Sedim. Petrol.* **55**, 769–773 (1985).

- Holt, R. M. & Powers, D. W. *US Department of Energy Report No. 90-051 1-90* (Carlsbad, New Mexico, 1990).
- Hovorka, S. *Sedimentology* **34**, 1029–1054 (1987).
- Bein, A., Hovorka, S. D., Fisher, R. S. & Roedder, E. *J. Sedim. Petrol.* **61**, 1–14 (1991).
- Casas, E. & Lowenstein, T. K. *J. Sedim. Petrol.* **59**, 724–739 (1989).
- Roberts, R. M., Beauheim, R. L. & Domski, P. S. *Sandia National Laboratories Report SAND98-2537* (Albuquerque, New Mexico, 1999).
- Stein, C. L. & Krumhansl, J. L. *Geochim. Cosmochim. Acta* **52**, 1037–1046 (1988).
- Lowenstein, T. K. *Geol. Soc. Am. Bull.* **100**, 592–608 (1988).
- O'Neil, J. R., Johnson, C. M., White, L. D. & Roedder, E. *Appl. Geochem.* **1**, 265–271 (1986).

Fisheries

Different behaviour of North and Irish Sea cod

Cod (*Gadus morhua*) are bottom-living, predatory fish of considerable commercial importance¹, but surprisingly little is known of what cod do for most of their time because it is difficult and costly to study the behaviour of fish at sea^{2,3}. Here we use electronic data-storage tags to investigate the behaviour of cod in the North Sea and in the Irish Sea and find that there are marked differences in the activity of fish in the two regions. This difference could be explained by dissimilar foraging ecology and may have implications for the future management of severely depleted cod stocks.

Evidence from acoustic⁴ and fishing⁵ surveys can be used to infer variations in behaviour patterns between different cod stocks, but behavioural data have never been collected simultaneously from individual fish from different stocks over timescales appropriate to fisheries management⁶. In March–April 1999, we tagged 78 cod (each over 50 cm long) with electronic data-storage devices: 58 were in the southern North Sea and 20 in the central Irish Sea. Twenty-two tags have so far been returned from fish caught in the North Sea, providing over 1,500 days of data, and four have been returned from the Irish Sea (over 750 days of data).

By using the depth record to estimate a cod's activity, we found that Irish Sea cod were extremely active at all times (dark regions in Fig. 1a), showing no discernible evidence of diurnal or seasonal patterns. North Sea cod were also very active during April and May, but in June all fish still at liberty (8 fish) showed a marked reduction in vertical movement, and in July they spent much or all of their time on the sea floor. Fish still free in August and September (5 fish) became active at night, and during October and November (2 fish) they returned to activity levels seen in the previous April and May (Fig. 1a). In a follow-up experiment in 2000, we monitored the position of North Sea cod implanted with

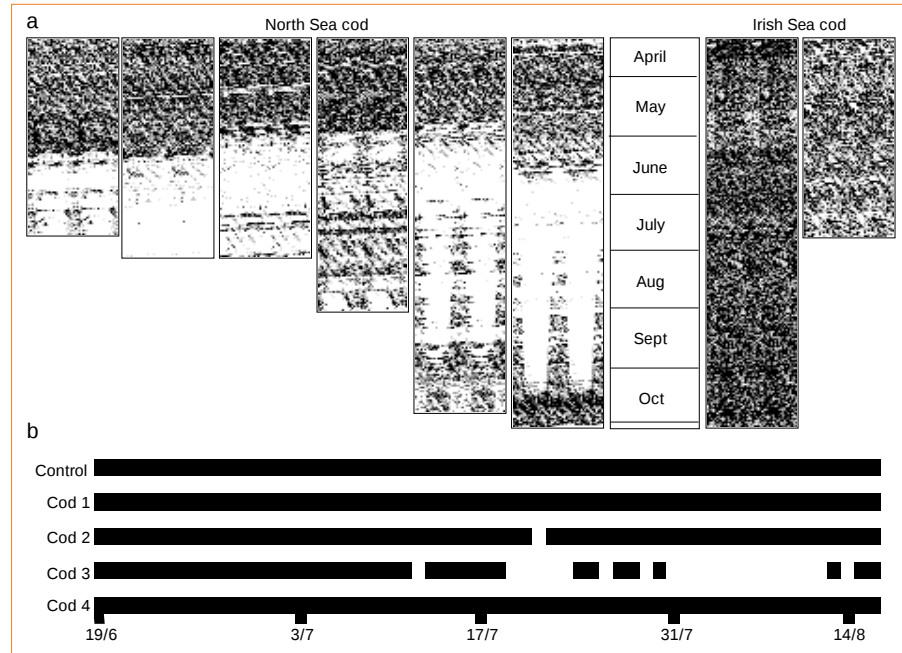


Figure 1 Cod behaviour in the North Sea and Irish Sea. **a**, Activity between April and November 1999. Active and inactive states of each individual were determined from the depth (measured every 10 min) record of its tag (Lotek Marine Technologies LTD100). When an individual was 'inactive' on the seabed, its tag recorded only the smooth changes in pressure resulting from the rise and fall of the tide. Individuals were classed as 'active' when vertical movements were more rapid or irregular than could be accounted for by the tide alone. For each hour of the day, summed hourly activity held a value between zero (white; inactive) and six (black; most active). Each fish's activity record is shown as a double-plot actogram, with data in 48-h periods along the x-axis and days of the year along the y-axis. **b**, Residence on summer feeding grounds in the North Sea in 2000. A moored listening station continuously monitored the presence (bars) or absence (no bar) of four individuals implanted with individually coded acoustic tags. The listening station had a detection range of 500 m. The control tag was located within 200 m of the listening station for the duration of the experiment.

individually coded acoustic tags and demonstrated that their range of movement during the summer months (mid-June to mid-August) was less than 1 km (Fig. 1b).

The technology is not yet available to measure the feeding behaviour of fish directly in the field. However, our results indicate that North Sea cod reduce their foraging movements during the summer months. This observation challenges the explicit assumptions of multispecies management models⁷ that cod are active and forage over substantial geographical areas. By contrast, Irish Sea cod appear to be active throughout the spring and summer. We propose that these differences in activity could be behavioural responses to variations in the distribution and abundance of prey species between sea basins, an idea we intend to test by using a tag equipped with a feeding sensor to investigate the feeding dynamics of free-ranging cod.

Low spawning-stock biomass, increased seawater temperatures and high fishing rates have put North Sea and Irish Sea cod stocks under great pressure^{8,9}. Special technical conservation measures designed to provide protection for cod have been specified in recovery plans announced in 2000 for the Irish Sea¹⁰ and recently for the North Sea¹¹. A significant component of these plans involves closing particular areas of sea to fishing for short periods of time.

Our findings show that the activity of cod varies through the year, and that seasonal patterns of activity differ between stocks. To our knowledge, this is the first time that the activity patterns of cod from different stocks have been described over extended periods at such a fine temporal and spatial scale. This understanding will improve our ability to predict the effectiveness of fisheries management by closed areas in different regions.

David Righton*, Julian Metcalfe*, Paul Connolly†

*Centre for Environment, Fisheries and Aquaculture Science, Lowestoft Laboratory, Lowestoft, Suffolk NR33 0HT, UK
e-mail: d.righton@cefas.co.uk

†Marine Fisheries Services Division, The Marine Institute, Abbotstown, Dublin 15, Eire

- Yearbook of Fishery Statistics **84** (FAO, Rome, 1997).
- Godø, O. R. & Michalsen, K. *Fisheries Res.* **48**, 127–140 (2000).
- Metcalfe, J. D. & Arnold, G. P. *Nature* **387**, 665–666 (1997).
- Wroblewski, J. S., Nolan, B. G., Rose, G. A. & deYoung, B. *Fisheries Res.* **45**, 51–59 (2000).
- Beamish, F. W. H. *J. Fish. Res. Bd Can.* **23**, 109–139 (1966).
- Arnold, G. P., Walker, M. G., Emerson, L. S. & Holford, B. H. *ICES J. Mar. Sci.* **51**, 207–232 (1994).
- Magnusson, K. G. *Rev. Fish. Biol. Fish.* **5**, 195–212 (1995).
- Report of the ICES Advisory Committee on Fisheries Management **236** (ICES, Copenhagen, 1999); <http://www.ices.dk>
- O'Brien, C. M., Fox, C. J., Planque, B. & Casey, J. *Nature* **404**, 142 (2000).
- Commission Regulation EC No. 304/2000. Official Journal. L 35/10 (2000).
- Commission Regulation EC No. 259/2001. Official Journal. L 39/7 (2001).

Closing of the Indonesian seaway as a precursor to east African aridification around 3–4 million years ago

Mark A. Cane* & Peter Molnar†

* Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 01964-8000, USA

† Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; and Department of Geological Sciences, Cooperative Institute for Research in Environmental Science, Campus Box 399, University of Colorado, Boulder, Colorado 80309, USA

Global climate change around 3–4 Myr ago is thought to have influenced the evolution of hominids, via the aridification of Africa, and may have been the precursor to Pleistocene glaciation about 2.75 Myr ago. Most explanations of these climatic events involve changes in circulation of the North Atlantic Ocean due to the closing of the Isthmus of Panama. Here we suggest, instead, that closure of the Indonesian seaway 3–4 Myr ago could be responsible for these climate changes, in particular the aridification of Africa. We use simple theory and results from an ocean circulation model to show that the northward displacement of New Guinea, about 5 Myr ago, may have switched the source of flow through Indonesia—from warm South Pacific to relatively cold North Pacific waters. This would have decreased sea surface temperatures in the Indian Ocean, leading to reduced rainfall over eastern Africa. We further suggest that the changes in the equatorial Pacific may have reduced atmospheric heat transport from the tropics to higher latitudes, stimulating global cooling and the eventual growth of ice sheets.

Eastern Africa used to be more humid than it is now, and Canada used to be devoid of ice sheets; New Guinea lay south of its present position, and the island of Halmahera (part of Indonesia; see Fig. 1) lay mostly below sea level. The collision of Australia with the Banda arc occurred about 3 Myr ago, when Timor began to emerge within what had been the eastward continuation of the Java trench^{1,2} (Fig. 1) and volcanism along the arc ceased³. Here we are concerned with the flow of water through the archipelago of Indonesia: that is, the Indonesian throughflow. The more important events that affected this throughflow, however, occurred farther north, associated with New Guinea's northward movement towards the Equator.

Hydrographic measurements show that almost all water currently passing through the Indonesian seaway derives from the North Pacific Ocean^{4,5} (Fig. 2). In the modern climate, heavy rainfall over the northern part of the western Pacific maintains fresher tropical waters north of the Equator than to the south. The configuration of landmasses, with more land to the north, forces this rainfall. Because this distribution of Asian landmass has changed little since Miocene time, we presume that the difference in salinity between western tropical Pacific waters north and south of the Equator has existed throughout this time. Being less saline than water at the same density in the southern Pacific, water in the northern Pacific is the colder, with a sharp temperature front at the Equator on surfaces of constant density (Fig. 2).

Warm southern Pacific water currently moves westward along the Equator, in the Southern Equatorial Current, along the north coast of New Guinea to the Halmahera eddy, just east of the island of Halmahera (Fig. 1), where it turns to flow eastward in the North Equatorial Countercurrent. We propose that when New Guinea lay farther south, and Halmahera was a smaller island, warm water from the South Pacific would have passed into the Indian Ocean, increasing sea surface temperatures (SSTs) there and precipitating a rainier climate in eastern Africa.

Closing of the Indonesian seaway

With a northward rate of about 70 km Myr⁻¹ (ref. 6) relative to Eurasia and Sundaland (the area that includes Sulawesi, Borneo, Java and Sumatra)—areas that have moved little with respect to the

Earth's spin axis in the past 20 Myr (ref. 7)—Australia and New Guinea lay 2°–3° south of their present positions at 3–5 Myr ago. Thus, a wide gap would have lain between New Guinea and Halmahera, which currently lies on the Pacific plate (or perhaps the Philippine Sea plate) and moves west relative to Sundaland.

Halmahera was a much smaller island 3–5 Myr ago than it is today. The western part of Halmahera consists of young volcanic rock. Ancient reefs on the eastern half of the island now lie at an elevation of 1,000 m (ref. 8), implying that its surface has risen by as much since about 5 Myr ago, consistent with an estimated ~60% east–west shortening of a once wider island⁹. Much, if not all, of Halmahera, which currently plays a major role in preventing water from the southern Pacific from joining the Indonesian throughflow¹⁰, apparently has emerged since 5 Myr ago.

The region between the 'Bird's Head' of New Guinea and Sulawesi deforms as rapidly as any on Earth at present^{11,12}. The Molucca Sea narrows at ~100 km Myr⁻¹; it also becomes shallower, as thick sediment in it becomes thrust atop itself¹³. Between the Bird's Head and Sulawesi, which converge at ~80 km Myr⁻¹, the various islands pile up one behind the other; this narrows and shallows the passages between the islands. But because relative motions among blocks have changed rapidly during the past 5 Myr, reconstructing the relative positions of the various islands of eastern Indonesia at ~5 Myr ago requires speculation that guarantees error, as indicated by differences among reconstructions^{14–18}. Nonetheless, the combination of a wider gap between Seram and the Bird's Head, which converge at ~50 mm yr⁻¹ (refs 11, 12), the recent rapid uplift of islands like Atauro along the eastern Banda arc¹⁹, and the absence of Timor¹², implies that only a few million years ago, much of the shallow sea floor between New Guinea and Sulawesi was deeper, and gaps between islands were wider. Thus, not only was New Guinea farther south, but there was also a seaway between the Pacific and Indian oceans as wide and deep as the Makassar Strait, but located to the east and south of that feature.

Environmental change in eastern Africa

Climate in eastern Africa has evolved from moist and warm to arid and perhaps slightly cooler over the past ~3–4 Myr (ref. 20, Fig. 3).

Except for the Fort Ternan site near the Equator in Kenya²¹, fossil fauna and flora suggest warm, moist environments, including rainforests, in eastern Africa before ~ 4 Myr ago^{22–24}. In central Ethiopia, wooded habitats (at 4.4 Myr ago²⁵), and even a rainforest environment, seem to have lingered until 3.4 Myr ago²⁶. Similarly, in the Omo region of southern Ethiopia, a humid environment

recorded by pollen at 4.1 Myr ago²⁶ lasted until 2.95 Myr ago, as implied by small mammals characteristic of that environment²⁷.

The tendency towards more arid environments apparently began earlier at lower latitudes. At Laetoli, Tanzania (Fig. 3), most fossil pollen dating from 3.7 Myr ago can be found in plants now living within 30 km of the region, suggesting little change in climate there

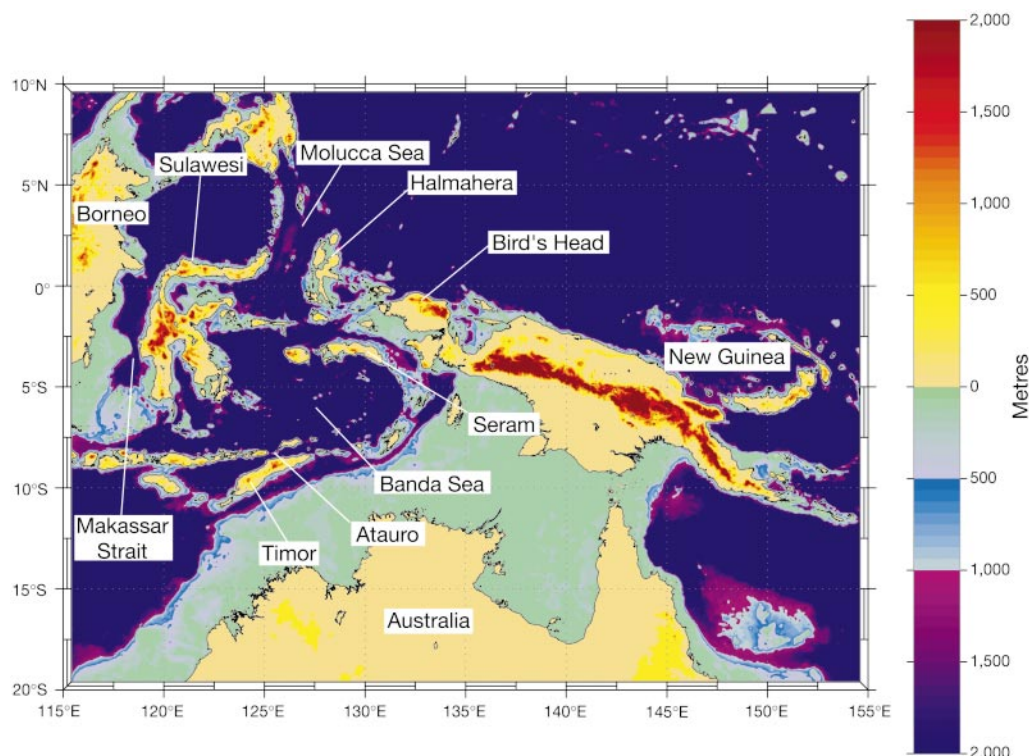


Figure 1 Map of the Indonesian-throughflow region. The main islands, crustal fragments, and other topographic features are labelled. Currently most Indonesian throughflow passes north of Sulawesi, then west of Sulawesi through the Makassar Strait, and finally across the Banda arc. The South Equatorial Current flows along the northern coast of New Guinea and then turns to the east at Halmahera in the Halmahera eddy. We suggest that at 3–5 Myr ago, Halmahera was a much smaller island with most of the area $\sim 1,000$ m

deeper, New Guinea was 2° – 3° farther south, and the mountains on the island were also much lower. In addition, Timor was under water, and the Java trench continued eastward and then northward into the Seram trough. Finally, the 'Bird's Head' lay 250–400 km east of its present position with respect to Borneo. Colour indicates elevation relative to mean sea level.

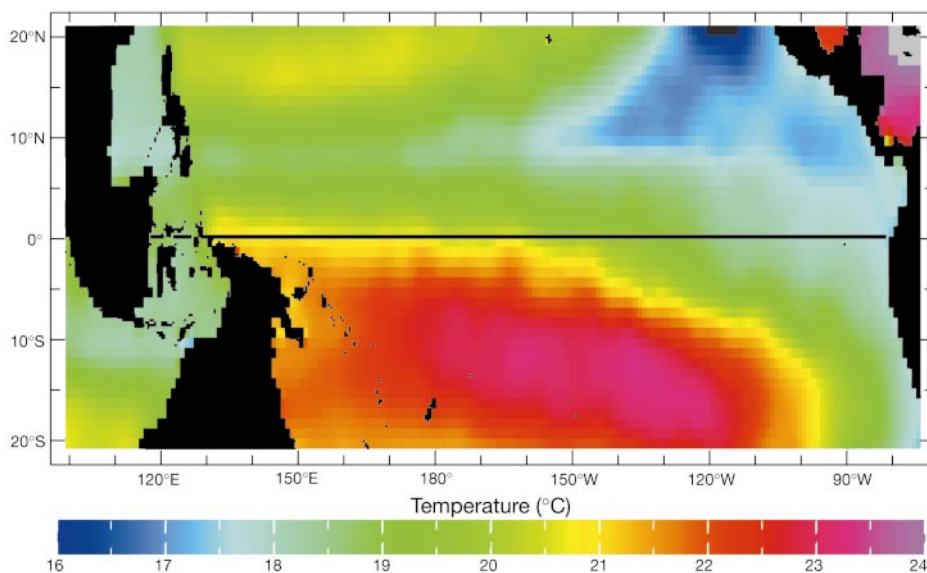


Figure 2 Map of the Pacific showing temperature at the $\sigma = 25.5$ ($1,025.5 \text{ kg m}^{-3}$) isopycnal, in the thermocline. We note the strong front at the Equator in the west. It is

apparent that water at this density in the Indonesian throughflow has a North Pacific source. Data from ref. 52.

since that time²¹. The aridification is best resolved in the Omo region (Fig. 3). Wesselman²⁷ showed that small mammals at 2.95 Myr ago included “species characteristic of the equatorial high forests of Central and West Africa,” but by a little before 2.52 Myr ago only one species (in fact, a “forest-edge taxon”) represented the tropical forest, and taxa typical of mesic woodlands dominated. The most marked shift in climate seems to have occurred between 2.4 and 2.5 Myr ago; by 2.34 Myr ago, riverine and forest taxa had diminished, and xeric species dominated²⁷. Then, by just after 2.32 Myr ago, forest taxa were gone, for the fossil record contains only “dry savannas/open savanna woodland” and “arid semiarid steppe” taxa²⁷. Farther north, in Ethiopia, evergreen bushland and montane forests at 3.3 and 2.9 Myr ago²⁸ gave way, a few tens of kilometres to the south, to grasslands by 2.5 Myr ago, though the climate there was not yet as arid as now²⁹.

The palaeontological and palynological data can neither resolve short-term climate changes at most localities nor define unambiguous spatial differences in the pattern of increasing aridification. Nowhere in Africa, however, does there seem to have been an environmental change as large or as abrupt as that implied by the increase in ice-rafted debris in the North Atlantic and North Pacific

at ~2.7 Myr ago, which presumably marks the onset of widespread glaciation in the Northern Hemisphere^{30,31}.

Indonesian throughflow and east African aridity

Upon entering the Indian Ocean, waters of the modern Indonesian throughflow travel due west across the Indian Ocean to the east coast of Africa³². It is likely that such simple zonal flow has persisted for at least the past 10 Myr, because conservation of vorticity dictates the zonal path across the ocean, and the relatively constant latitude at which the water enters the Indian Ocean (~10°S) sets the vorticity of that water.

Rodgers *et al.*³³ used a general circulation model (GCM) of the ocean to calculate the difference between an approximation to present-day conditions and a case with the northern edge of New Guinea 3° south of its present position. It showed temperatures at 100 m depth across the central Indian Ocean to be 2°C warmer for the more southerly position (Fig. 4). The difference reaches a maximum of 3°C in a small region in the central Indian Ocean. Thus, a few million years ago, the central Indian Ocean may have been a few degrees warmer than it is today.

Observational studies of the modern climate have shown a link between east African rainfall and Indian Ocean SST, with warmer temperatures associated with more rain^{34–36}. Several papers^{34,35} offer explanations for the observed associations in which warm SST anomalies induce changes in the zonal (Walker-like) atmospheric circulation over the Indian Ocean, reducing the subsidence over east Africa and thus increasing rainfall there. The link between Indian Ocean SST and rain in east Africa was exhibited in the severe flooding that occurred during the 1997–98 El Niño event. Indeed, there is a well studied positive correlation between rainfall in eastern Africa and the warm (El Niño) phase of the El Niño/Southern Oscillation (ENSO) cycle³⁷, and an even stronger correlation between ENSO and Indian Ocean SST. Because Indian Ocean SST anomalies are so hard to separate from other ENSO changes, data alone are insufficient to decide if the Indian Ocean SST anomalies cause the African rainfall variations.

An atmospheric GCM forced by SSTs³⁶ shows that warmer temperatures in the eastern Pacific, by themselves, slightly reduce calculated rainfall over Africa. A warmer central Indian Ocean alone, however, yields calculated rainfall over eastern Africa that is even greater than the full ENSO response, and an increase in temperature of only 0.5°C can induce the observed high rainfall in Africa during El Niño years in the GCM calculations^{36,38}. Goddard and Graham³⁶ showed that a change in the circulation of the atmosphere acting on the mean moisture field provides the main

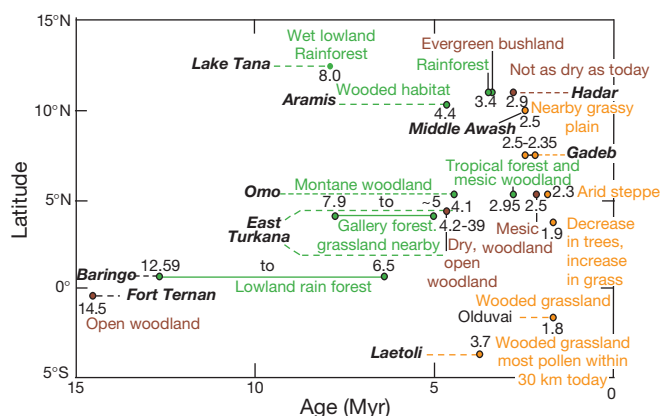


Figure 3 Palaeo-environments in eastern Africa as a function of time and latitude since 15 Myr ago. Green, humid environments; brown, more arid woodlands; and tan, more arid grasslands and deserts. Data sources as follows: Lake Tana²⁴, Hadar^{26,28}, Aramis²⁵, Middle Awash²⁹, Gadeb⁶³, Omo^{26,27}, East Turkana^{21,23,54}, Baringo²², Fort Ternan, Olduvai and Laetoli²¹. Numbers indicate ages in Myr.

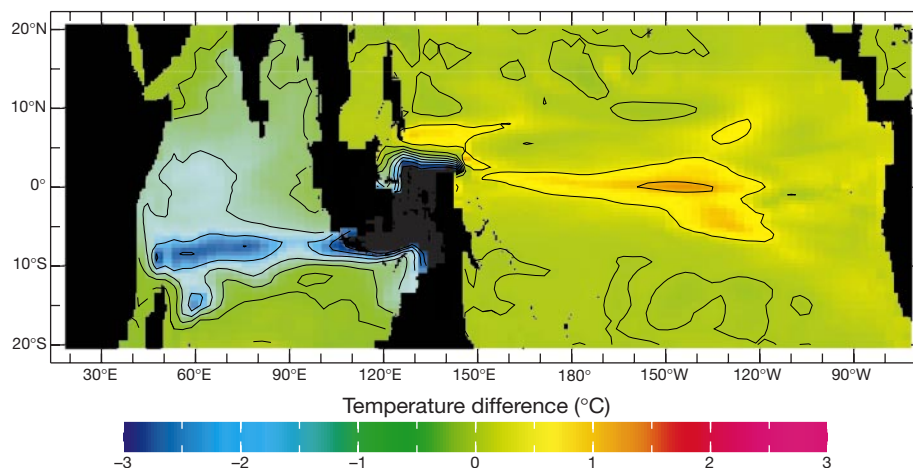


Figure 4 The difference in temperatures at 100 m depth between the two ocean GCM runs of ref. 33. The sole difference between conditions for the runs is that in one the northern tip of New Guinea–Halmahera is at 2°N, and in the other it is at 3°S. We suggest

that the pattern shown—the difference of the former minus the latter—approximates the difference between that at present minus that at 4 Myr ago.

mechanism by which rainfall increases in the GCM calculations. This model result is corroborated by their study of reanalysis data, and is essentially the same mechanism suggested in earlier observational studies^{34,35}. The atmospheric GCM studies^{36,38} leave little doubt that Indian Ocean SSTs are the controlling causal factor affecting variations in rainfall over east Africa.

Indonesian throughflow and the position of New Guinea

The results of numerical-model experiments³³ (Fig. 4)—which we use as evidence that the Indian Ocean would warm as the western Pacific barrier ‘island’ of Australia–New Guinea moved northward—may be thought to depend too strongly on the peculiarities of the model, or the modern wind field used in these experiments. We present here a more general result. Because most of the Indonesian throughflow occurs in the upper part of the water column⁵, and because SSTs influence the atmosphere most, we consider the upper ocean through the thermocline. It is therefore appropriate to use a theory for baroclinic flow, with parameter settings for the gravest baroclinic (largest vertical wavelength) mode. Because east–west variations in flow span long distances (long wavelengths), islands can be treated as one-dimensional features for which only their northern and southern limits matter³⁹. This simplification allows the production of analytical theories for such flow; such theories have been worked out independently by two groups^{40,41}, and yield identical results for the problem we consider. Insofar as the theories are valid, only two aspects of the complex geometry of the region matter for the Indonesian throughflow; the northern latitude (b) of the Halmahera–New Guinea–Australia barrier on the Pacific Ocean side; and the southern latitude (g) of the collection of islands comprising Java, Sulawesi, Borneo, and all of Asia that form a barrier on the Indian Ocean side.

Following the approach summarized in the Methods section, and to be presented in more detail elsewhere, we obtain an estimate of the amplitude of the Indonesian throughflow, θ , for a unit mass flux (a δ -function) impinging on the western boundary of the Pacific at latitude c . Because Sundaland was essentially in place before 10–20 Myr ago, and the southern limit of Java has moved little during this period, we fix g at 10°S. With such a value, θ depends strongly on b , and in particular on which hemisphere the northern tip of Australia–New Guinea (with or without Halmahera) lies (Fig. 5). When b is south of the Equator, the Indonesian passage is effectively blocked for incident flux north of the Equator; only ~2% of Northern Hemisphere ($c > 0$) water passes through, but ~32% of the flux south of the Equator ($c < 0$) passes through. The situation reverses when b moves north of the Equator; only ~2% of South Pacific westward flux ($c < 0$) enters the Indian Ocean, but ~25% of northern waters ($c > 0$) form the Indonesian throughflow. As either b or c passes near the Equator (within one radius of deformation, which is ~3° of latitude for the gravest baroclinic mode) there is a rapid transition from one regime to the other (Fig. 5).

This solution has a simple physical interpretation, because in the low-frequency limit, large-scale motions in the open ocean can be described by waves—westward-propagating Rossby waves, and equatorially trapped, eastward-propagating, Kelvin waves. Whereas westward propagation can occur at all latitudes, the eastward return propagation occurs only near the Equator. When the westward-propagating jet meets a western boundary, like that of the Australia–New Guinea island, it becomes a current trapped to the boundary. The Coriolis effect forces the boundary current to propagate towards the Equator. If the northern extent of the island lies south of the Equator ($b < 0$), then when the current reaches the northern end, most of the boundary current turns west past the island. If the island reaches into the Northern Hemisphere ($b > 0$), however, the boundary current turns eastward and propagates as an

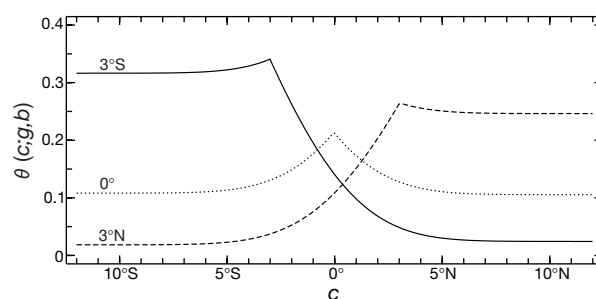


Figure 5 Plots of the Indonesian throughflow $\theta(c; g, b)$ for a δ -function zonal flow incident on the western boundary of the Pacific at latitude c . The plots are for three values of b , the latitude of the northern tip of the western Pacific boundary now ending at Halmahera: $b = 3^\circ\text{S}$, 0° and 3°N . In all cases, the latitude of the Indian Ocean boundary formed by Java to Timor is $g = 10^\circ\text{S}$. Calculations are according to the analytic theory described in Methods.

equatorial Kelvin wave. Correspondingly, when a jet north of the island reaches the western island (Sundaland–Eurasia), it becomes a western boundary current travelling to the Equator, where it turns east as Kelvin waves. If $b > 0$, reflection of the Kelvin waves from the west side of the Australia–New Guinea island traps most of the flow in the Indonesian seas, creating an Indonesian throughflow. If instead $b < 0$, the Kelvin waves will propagate eastward unobstructed, transporting a large fraction of the jet back to the Pacific rather than letting it pass into the Indian Ocean.

The function $\theta(c; g, b)$ plotted in Fig. 5 is a Green’s function; for a general mass flux $u(y)$ at the western boundary of the Pacific, the magnitude of the Indonesian throughflow is given by $\int_{-\infty}^{\infty} \theta(y; g, b) u(y) dy$. The incident mass flux $u(y)$ depends on the wind stress over the Pacific, as well as on global thermohaline forcing. Although little data constrain what these forcing functions were several million years ago, because of the simple structure of θ we may conclude that the movement of b from south to north must have changed the dominant source of the Indonesian throughflow from the South Pacific to the North Pacific. With $b = 3^\circ\text{S}$, virtually all of the water flowing into the Indian Ocean would have been from the south, whereas with $b = 3^\circ\text{N}$, almost all of the water would be derived from the north. In the modern ocean, where $b \approx 3^\circ\text{N}$, compelling evidence shows that the Indonesian throughflow comprises mostly northern water^{4,5}. The total transport would change little as b progressed from 3°S to 3°N , probably decreasing slightly. These statements should hold unless $u(y)$ were so skewed that the total flux from one hemisphere were at least 10 times the other, but it is hard to imagine how this extreme state could arise in a configuration so much like that of the present day.

As New Guinea–Australia moved northward, the colder waters from the north replaced the warmer southern water in the Indonesian throughflow, cooling the thermocline in the Indian Ocean. Consequently, SSTs cooled throughout the Indian Ocean, especially in upwelling regions, such as the Somali Current north of the Equator along the African coast. We presume that the atmospheric response was like that in the modern climate, and so would reduce rainfall over eastern Africa. While we cannot assert categorically that this is true, it is the most plausible possibility: the proposed physical mechanism applies to present-day conditions, and the geometry of the Indian Ocean was very much as it is today. The northward movement of New Guinea–Australia would have caused the drying of east Africa by switching the source waters for the Indonesian throughflow from the warm south Pacific to the cold north, thereby cooling the Indian Ocean and reducing rainfall in east Africa.

New Guinea’s movement and global climate change

The onset of widespread glaciation in the Northern Hemisphere is

notable, not only for how it affected temperature and precipitation over large areas, but also for its abrupt beginning at ~ 2.75 Myr ago^{30,31}. Yet the oxygen-isotope content of benthic microorganisms, the dominant proxy for high-latitude temperature, show a gradual increase in $\delta^{18}\text{O}$ beginning at ~ 4 Myr ago⁴², with the most obvious change since then being an increase in the amplitude of variations. The absence of an abrupt change in the $\delta^{18}\text{O}$ record accords with a relatively slow geologic process affecting high-latitude climate gradually, not abruptly. (Glaciation may have set in abruptly when the decreasing temperature crossed the threshold for freezing water.) The closure of the Indonesian seaway provides such a gradual change at the right time. Given our prejudice for the tropical Pacific as a key to global climate change⁴³, we consider, necessarily only qualitatively, how the northward movement of New Guinea might affect global climate by changing the tropical Pacific.

In agreement with observations⁴⁴, our analytic solution (Fig. 5) indicates that with the present configuration the water turned eastward along the Equator in the Pacific comes from the south. Conversely, with the opening south of the Equator the source would be the colder water to the north, which agrees with GCM calculations³³ showing that the thermocline thins more in the western than in the eastern Pacific, and that the greatest surface cooling occurs over the central equatorial Pacific (Fig. 4). In these calculations, the gentler slope of the thermocline is maintained by specifying present-day easterly winds. With New Guinea both farther south and apparently lower, for the present-day high elevations of New Guinea reportedly formed since ~ 5 Myr ago^{45,46}, atmospheric convection over the western Pacific should have been weaker at ~ 3 – 5 Myr ago than now. The resulting weaker Walker circulation implies weaker easterlies over the equatorial Pacific, which would relax the slope of the thermocline further, warming the east and cooling the west⁴⁷. Changes in both the distributions of microorganisms living near the thermocline and oxygen isotopes of surface-dwelling (planktonic) foraminifers corroborate such a change in the western equatorial Pacific since 3–4 Myr ago⁴⁸. In a state much like a permanent El Niño, the ocean and especially the atmosphere would transport more heat from the tropics to higher latitudes. If the global anomaly pattern before 3 Myr ago resembled what currently occurs during El Niño, the absence of large continental ice sheets in Canada before 3 Myr ago follows from the greater heat transport to Canada⁴⁹.

We have suggested that the aridification of east Africa at ~ 4 Myr ago was caused by the northward movement of New Guinea–Australia and the emergence of much of Halmahera. This suggestion could be tested with sediment cores from the Indian Ocean, particularly with cores from 10°S , where we expect temperatures to have decreased. The possibility that the concomitant changes in the tropical Pacific played a major role in bringing about the onset of glaciation, if yet more speculative, could also be tested with equatorial Pacific sediment cores, and with data from regions known to have strong ENSO teleconnections in today's climate. $\square$

Methods

We use the linear shallow-water equations on an equatorial β -plane to approximate the baroclinic flow near the Equator⁵⁰. In the approximate form appropriate to low-frequency motions (including the steady state)⁵¹, the equations permit three classes of free motions, which must be described mathematically by orthogonal functions: (1) Rossby waves with long zonal wavelengths and standing meridional structures, which carry energy westward; (2) equatorial Kelvin waves, with the meridional structure $u(y) = h(y) = \psi_0(y) = \pi^{-1/4} \exp(-y^2/2)$, where u is the eastward velocity and h is the perturbation pressure or, equivalently, the perturbation height. All quantities are rendered dimensionless in the canonical way for equatorial motions^{50,51}. (3) Western boundary currents, trapped along that boundary.

The low-frequency, long-wavelength Kelvin and Rossby modes obey a geostrophic balance between the meridional pressure gradient and the Coriolis force:

$$yu + \frac{\partial h}{\partial y} = 0 \quad (1)$$

Flows incident on the western boundary are composed of modes that carry energy

westward: the long-wavelength Rossby waves. The reflection must carry energy eastward, and hence the reflected motion must consist of Kelvin waves. For a boundary devoid of passages, all mass flux incident on the boundary returns in Kelvin waves⁵¹, which determines the amplitude of the Kelvin mode. Western boundary currents connect the fluid circuits between the incident Rossby motions and the reflected Kelvin mode.

The western Pacific boundary has openings, allowing some of the fluid to be transmitted through to the Indian Ocean. By using geostrophy, equation (1), the condition that no Kelvin wave can be transmitted west of the boundary (namely, equation (4)), and the fact that all motions must be of types (1)–(3), the reflected and transmitted components of the flow may be determined^{40,41}.

The mass flux αT reflected and carried eastward by Kelvin waves, when a zonal current $u(y)$ and its geostrophically balanced pressure field $h(y)$ are incident on the western boundary of the Pacific, is given by:

$$\alpha T = - \frac{\alpha h(b) + \alpha b \int_{-\infty}^b u(y) dy + \Gamma(g) \int_{-\infty}^b u(y) dy}{\Gamma(g) + \Gamma(-b)} \quad (2)$$

(which corrects a typographical error in equation (41) of ref. 41) where T is the amplitude of the reflected Kelvin waves, $\alpha = \int_{-\infty}^b \psi_0(y) dy = \sqrt{2\pi}^{1/4}$ is the (dimensionless) mass flux in a Kelvin wave of unit amplitude, g is the latitude of the southern boundary of Eurasia–Sundaland, and b is the latitude of the northern boundary of Australia–New Guinea (with and without Halmahera). Latitudes are rendered dimensionless using the radius of deformation for the gravest baroclinic mode (333 km). Also in equation (2),

$$\Gamma(a) = \psi_0(a) - a \int_a^\infty \psi_0(y) dy \quad (3)$$

The minus sign in equation (2) accounts for the change of sign upon reflection.

If the incident mass flux u is a δ -function at latitude c , $u(y) = \delta(y-c)$, then by geostrophy, equation (1), $h(y) = h_-$ for $y < c$, and $h(y) = h_- - c$ for $y > c$, where h_- is a constant determined by the condition that $u(h)$ be orthogonal to the Kelvin wave:

$$\int_{-\infty}^\infty (u + h) \psi_0 dy = 0 \quad (4)$$

which yields $\alpha h_- = -\Gamma(c)$. Substituting the expressions for u and h into equation (2) we obtain:

$$\alpha T = - \frac{\Gamma(g) - \Gamma(c) + \alpha \max(0, b - c)}{\Gamma(g) + \Gamma(-b)} \quad (5)$$

The Indonesian throughflow due to a westward jet of unit amplitude, which equals incident minus reflected flux $\theta = 1 + \alpha T$, is plotted in Fig. 5.

Received 15 September 2000; accepted 7 March 2001.

- Audley-Charles, M. G. in *Thrust and Nappe Tectonics* 407–416 (Geological Society, London, 1981).
- Veevers, J. J., Falvey, D. A. & Robins, S. Timor Trough and Australia: facies show topographic wave migrated 80 km during the past 3 m.y. *Tectonophysics* **45**, 217–227 (1978).
- Abbott, M. J. & Chamalaun, F. H. in *The Geology and Tectonics of Eastern Indonesia* (eds Barber, A. J. & Wiryosujono, S.) 253–268 (Spec. Publ. No. 2, Geological Research and Development Centre, Bandung, 1981).
- Gordon, A. L. & Fine, R. A. Pathways of water between the Pacific and Indian oceans in the Indonesian seas. *Nature* **379**, 146–149 (1996).
- Gordon, A. L., Susanto, R. D. & Ffield, A. Throughflow within the Makassar strait. *Geophys. Res. Lett.* **26**, 3325–3328 (1999).
- DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Effects of a recent revisions to geomagnetic reversal timescale on estimates of current plate motions. *Geophys. Res. Lett.* **21**, 2191–2194 (1994).
- Besse, J. & Courtillot, V. Revised and synthetic apparent polar wander paths of the African, Eurasian, North American and Indian plates, and true polar wander since 200 Ma. *J. Geophys. Res.* **96**, 4029–2050 (1991).
- Hall, R., Audley-Charles, M. G., Banner, F. T., Hidayat, S. & Tobing, S. L. Late Paleogene–Quaternary geology of Halmahera, eastern Indonesia: initiation of a volcanic island arc. *J. Geol. Soc. Lond* **145**, 577–590 (1988).
- Nichols, G. J. & Hall, R. Basin formation and Neogene sedimentation in a backarc setting, Halmahera, eastern Indonesia. *Mar. Petrol. Geol.* **8**, 50–61 (1991).
- Morey, S. L., Shriver, J. F. & O'Brien, J. J. The effects of Halmahera on the Indonesian Throughflow. *J. Geophys. Res.* **104**, 23281–23296 (1999).
- Puntodewo, S. S. O. et al. GPS measurements of crustal deformation within the Pacific–Australia plate boundary in Irian Jaya, Indonesia. *Tectonophysics* **237**, 141–153 (1994).
- Rangin, C. et al. Plate convergence measured by GPS across the Sundaland/Philippine Sea plate deformed boundary: the Philippines and eastern Indonesia. *Geophys. J. Int.* **139**, 296–316 (1999).
- Silver, E. A. & Moore, J. C. The northern Molucca Sea collision zone, Indonesia. *J. Geophys. Res.* **83**, 1681–1691 (1978).
- Charlton, T. R. A plate tectonic model of the eastern Indonesian collision zone. *Nature* **319**, 394–396 (1986).
- Daly, M. C., Cooper, M. A., Wilson, I., Smith, D. G. & Hooper, B. G. D. Cenozoic plate tectonics and basin evolution in Indonesia. *Mar. Petrol. Geol.* **8**, 2–21 (1991).
- Hall, R. in *Tectonic Evolution of Southeast Asia* (eds Hall, R. & Blundell, D.) 153–184 (Spec. Publ. 106, Geological Society, London, 1996).
- Packham, G. in *Tectonic Evolution of Southeast Asia* (eds Hall, R. & Blundell, D.) 123–152 (Spec. Publ. 106, Geological Society, London, 1996).
- Rangin, C., Jolivet, L., Pubellier, M. & the Tethys Pacific working group. A simple model for the tectonic evolution of southeast Asia and Indonesia region for the past 43 m.y. *Bull. Soc. Geol. Fr.* **VI**, 889–905 (1990).
- Chappell, J. & Veeh, H. H. Late Quaternary tectonic movements and sea-level changes at Timor and Atauro Island. *Geol. Soc. Am. Bull.* **89**, 356–368 (1978).

20. deMenocal, P. B. Plio-Pleistocene African climate. *Science* **270**, 53–59 (1995).
21. Bonnefille, R. in *The Evolution of the East Asia Environment* Vol. II, *Palaeobotany, Palaeozoology, and Palaeoanthropology* (ed. Whyte, R. O.) 579–612 (Univ. Hong Kong, 1984).
22. Hill, A. in *Paleoclimate and Evolution with Emphasis on Human Origins* (eds Vrba, E. S., Denton, G. H., Partridge, T. C. & Burckle, L. H.) 178–193 (Yale Univ. Press, 1995).
23. Leakey, M. G. *et al.* Lothagam: A record of faunal change in the late Miocene of East Africa. *J. Vert. Paleontol.* **16**, 556–570 (1996).
24. Yemane, K., Bonnefille, R. & Faure, H. Paleoclimatic and tectonic implications of Neogene microflora for the northwestern Ethiopian highlands. *Nature* **318**, 653–656 (1985).
25. WoldeGabriel, G. *et al.* Ecological and temporal placement of the early Pliocene hominids at Aramis, Ethiopia. *Nature* **371**, 330–333 (1994).
26. Bonnefille, R. *Paleoclimate and Evolution with Emphasis on Human Origins* (eds Vrba, E. S., Denton, G. H., Partridge, T. C. & Burckle, L. H.) 299–310 (Yale Univ. Press, 1995).
27. Wesselman, H. B. *Paleoclimate and Evolution with Emphasis on Human Origins* (eds Vrba, E. S., Denton, G. H., Partridge, T. C. & Burckle, L. H.) 356–368 (Yale Univ. Press, New Haven, CT, 1995).
28. Bonnefille, R., Vincens, A. & Buchet, G. Palynology, stratigraphy and palaeoenvironment of a Pliocene hominid site (2.9–3.3 M. Y.) at Hadar, Ethiopia. *Palaeoclim. Palaeogeogr. Palaeoecol.* **60**, 249–281 (1987).
29. de Heinzelin, J. *et al.* Environment and behavior of 2.5-million-year-old Bouri hominids. *Science* **284**, 625–629 (1999).
30. Haug, G. H., Sigman, D. M., Tiedemann, R., Pedersen, T. F. & Sarinthein, M. Onset of permanent stratification in the subarctic Pacific Ocean. *Nature* **401**, 779–782 (1999).
31. Shackleton, N. J. *et al.* Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the North Atlantic region. *Nature* **307**, 620–623 (1984).
32. Gordon, A. L. Inter-ocean exchange of thermocline water. *J. Geophys. Res.* **91**, 5037–5046 (1986).
33. Rodgers, K. B., Latif, M. & Legutke, S. Sensitivity of equatorial Pacific and Indian Ocean watermasses to the position of the Indonesian throughflow. *Geophys. Res. Lett.* **27**, 2941–2945 (2000).
34. Hastenrath, S., Nicklis, A. & Greischar, L. Atmospheric-hydrospheric mechanisms of climate anomalies in the western equatorial Indian Ocean. *J. Geophys. Res.* **98**, 20219–20235 (1993).
35. Reverdin, G., Cadet, D. L. & Gutzler, D. Interannual displacements of convection and surface circulation over the equatorial Indian Ocean. *Q. J. R. Meteorol. Soc.* **112**, 43–67 (1986).
36. Goddard, L. & Graham, N. E. Importance of the Indian Ocean for simulating rainfall anomalies over eastern and southern Africa. *J. Geophys. Res.* **104**, 19099–19116 (1999).
37. Indeje, M., Semazzi, F. H. M. & Ogallo, L. J. ENSO signals in East African rainfall seasons. *Int. J. Climatol.* **20**, 19–46 (2000).
38. Latif, M., Dommengot, D., Dima, D. & Grotzner, A. The role of Indian Ocean sea surface temperature in forcing East African rainfall anomalies during December/January 1997/98. *J. Clim.* **12**, 3497–3504 (1999).
39. Cane, M. A. & du Penhoat, Y. The effect of islands on low frequency equatorial motions. *J. Mar. Res.* **40**, 937–962 (1982).
40. Clarke, A. J. On the reflection and transmission of low-frequency energy at the irregular western Pacific Ocean boundary. *J. Geophys. Res.* **96**, 3289–3305 (1991).
41. du Penhoat, Y. & Cane, M. A. Effect of low-latitude western boundary gaps on the reflection of equatorial motions. *J. Geophys. Res.* **96**, 3307–3322 (1991).
42. Haug, G. H. & Tiedemann, R. Effect of the formation of the Isthmus of Panama on Atlantic thermohaline circulation. *Nature* **393**, 673–676 (1998).
43. Cane, M. A. A role for the tropical Pacific. *Science* **282**, 59–60 (1998).
44. Johnson, J. C. & McPhaden, M. J. Interior pycnocline flow from the subtropical to the equatorial Pacific Ocean. *J. Phys. Oceanogr.* **29**, 3073–3089 (1999).
45. Dow, D. B. A *Geological Synthesis of Papua New Guinea* (Geol. Geophys. Bull. 201, Bureau of Mineral Resources, Australian Govt Publication Service, Canberra, 1977).
46. Smith, I. E. & Davies, H. L. *Geology of the Southeast Papuan Mainland* (Geol. Geophys. Bull. 165, Bureau of Mineral Resources, Australian Govt Publication Service, Canberra, 1976).
47. Fedorov, A. V. & Philander, S. G. Is El Niño changing? *Science* **288**, 1997–2002 (2000).
48. Chaisson, W. P. & Ravelo, A. C. Pliocene development of the east-west hydrographic gradient in the equatorial Pacific. *Paleoceanography* **15**, 497–505 (2000).
49. Trenberth, K. E. *et al.* Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.* **103**, 14291–14324 (1998).
50. Moore, D. W. & Philander, S. G. H. in *The Sea* Vol. 6 (eds Goldberg, E. D., McCave, I. N., O'Brien, J. J. & Steele, J. H.) 319–361 (Wiley, New York, 1977).
51. Cane, M. A. & Sarachik, E. S. Forced baroclinic ocean motion II: The equatorial unbounded case. *J. Mar. Res.* **35**, 395–432 (1977).
52. Levitus, S. & Boyer, T. P. *World Ocean Atlas 1994* Vol. 4, *Temperature* (US Dept of Commerce, Washington DC, 1994).
53. Bonnefille, R. Evidence for a cooler and drier climate in the Ethiopian uplands toward 2.5 Myr ago. *Nature* **303**, 487–491 (1983).
54. Leakey, M. G., Feibel, C. S., MacDougall, I. & Walker, A. New four-million-year-old hominid species from Kanapoi and Allia Bay, Kenya. *Nature* **376**, 565–571 (1995).

Acknowledgements

We thank K. Rodgers for providing the data for Fig. 4 in advance of publication. We also thank A. L. Gordon, P. deMenocal, H. Davies, L. Goddard, J. S. Godfrey, G. Krahmann, R. McCaffrey, N. Naik, S. G. Philander, M. Pubellier and K. Rodgers for help with figures, guidance and timely inspiration. M.A.C. was supported in part by the National Science Foundation.

Correspondence and requests for materials should be addressed to M.A.C (e-mail: mcane@ldeo.columbia.edu).

Evidence for planet engulfment by the star HD82943

G. Israelian*, N. C. Santos†, M. Mayor† & R. Rebolo*‡

* Instituto de Astrofísica de Canarias, E-38200 La Laguna, Tenerife, Spain

† Observatoire de Genève, 51 ch. des Maillettes, CH-1290 Sauverny, Switzerland

‡ Consejo Superior de Investigaciones Científicas, 28006, Madrid, Spain

Current models^{1,2} of the evolution of the known extrasolar planetary systems need to incorporate orbital migration and/or gravitational interactions among giant planets to explain the presence of large bodies close to their parent stars. These processes could also lead to planets being ingested by their parent stars, which would alter the relative abundances of elements heavier than helium in the stellar atmospheres. In particular, the abundance of the rare ⁶Li isotope, which is normally destroyed in the early evolution of solar-type stars³ but preserved intact in the atmospheres of giant planets, would be boosted substantially. ⁶Li has not hitherto been observed reliably in a metal-rich star^{4,5}, where metallicity refers to the total abundance of elements

heavier than helium. Here we report the discovery of ⁶Li in the atmosphere of the metal-rich solar-type star HD82943, which is known to have an orbiting giant planet. The presence of ⁶Li can probably be interpreted as evidence for a planet (or planets) having been engulfed by the parent star.

A planet with a minimum mass of 2.2M_J (where M_J is the mass of Jupiter) in a long-period orbit around the solar-type star HD82943 has been discovered⁶. The study of the Li line in this star using high-quality ($R = \lambda/\Delta\lambda \sim 50,000$) spectra⁷ showed an asymmetry in the red wing, which is strongly suggestive of the presence of a significant amount of ⁶Li. Using ⁶Li as a possible indicator of a planet-engulfing scenario, we obtained three high-resolution ($R \sim 110,000$) spectra of this star on 7 June 2000 with the UVES spectrograph on the 8 m VLT Kueyen. The comparison star, HD91889, and an early-type fast rotator, HD101431, were also observed using the same instrumental setup. The individual spectra are in very good agreement and, after co-adding, provided spectra in the Li line region with a signal-to-noise (S/N) ratio of around 500. The wavelength calibration was done with a thorium–argon lamp, resulting in a typical root mean square (r.m.s.) deviation from the seventh order polynomial fit of 4 mÅ. According to recent work⁷, the stellar parameters of HD82943 are $T_{\text{eff}} = 6,010 \pm 50$ K, $\log g = 4.62 \pm 0.20$, $[\text{Fe}/\text{H}] = +0.32 \pm 0.06$ (abundances for elements are

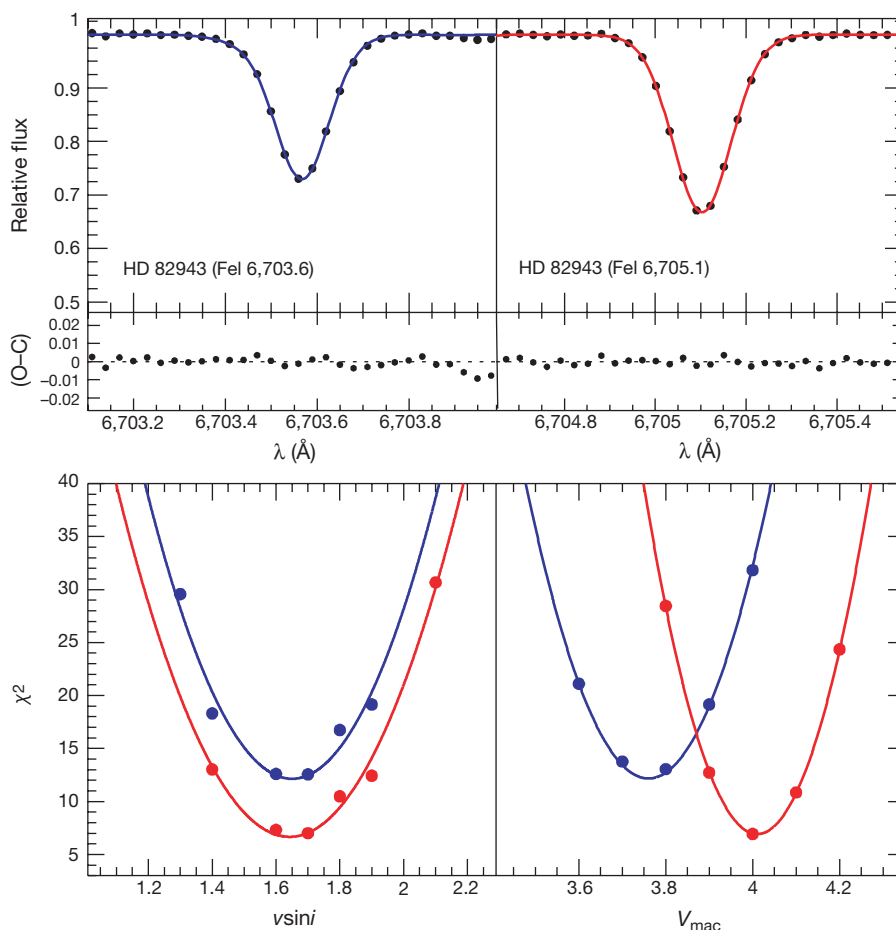


Figure 1 Observed and computed profiles of two reference iron lines (top) and results of the χ^2 tests (bottom). Filled circles in upper panels, observations; continuous lines, synthetic spectra derived from the χ^2 analysis. The χ^2 tests were performed considering the wavelengths, abundances, microturbulent velocities and broadening functions of the Fe lines as free parameters. Small adjustments of the continuum level were allowed to optimize the fits. The synthetic spectrum was convolved with a gaussian function representing the instrumental profile; various combinations of different broadening functions were applied; and the best fit to the observed profiles of the Fe lines was achieved after the synthetic spectra were convolved with both a radial–tangential

macroturbulence (V_{mac}) and a rotational profile ($vsini$) employed in our spectrum synthesis code. Blue and red colours correspond to the Fe I lines at 6,703.56 and 6,705.1 Å, respectively. The residuals ($O - C$) of the observations after subtraction of the synthetic spectra are also shown. The broadening parameters of the two Fe lines provide a mean $vsini = 1.65 \pm 0.05$ km s^{−1} (1σ error) and $V_{\text{mac}} = 3.9 \pm 0.2$ km s^{−1} for HD82943. Similar-quality results were achieved for HD91889 ($vsini = 2.75 \pm 0.1$ km s^{−1}, $V_{\text{mac}} = 6.2 \pm 0.1$ km s^{−1}). However, given the low $vsini$ of HD82943, it is possible to obtain fits using a simple gaussian function that are almost as good as those shown in the upper panels. Our results are not affected by these changes.

$[X/H] = \log[N(X)/N(H)]_{\text{star}} - \log[N(X)/N(H)]_{\text{Sun}}$, where $N(X)$ is the number density of atoms and microturbulence $\xi = 1.08 \pm 0.10 \text{ km s}^{-1}$. Using the same spectral-synthesis techniques⁷ we obtained $T_{\text{eff}} = 6,070 \pm 50 \text{ K}$, $\log g = 4.41 \pm 0.20$, $[\text{Fe}/\text{H}] = -0.23 \pm 0.05$ and $\xi = 1.52 \pm 0.10 \text{ km s}^{-1}$ for the comparison star HD91889.

We used a standard profile-fitting procedure^{8–10} and set the line-broadening parameters using two Fe I lines at 6,703.56 and 6,705.1 Å (Fig. 1). The atomic parameters of the Fe I lines used in our analysis were slightly modified from ref. 9 to reproduce the high-resolution solar atlas¹¹ with the solar abundance¹² of Fe. This procedure was repeated for the CN lines located in the Li line region¹³. Precise measurements of the wavelengths and the oscillator strengths of the Li lines are available⁸. Finally, we used synthetic spectra to select several spectral windows free from lines, to normalize the continuum and to estimate the S/N ratio. Average broadening parameters derived from the Fe lines were used to determine the Li isotopic ratio from a χ^2 analysis using the profile-fitting method^{8–10}. The total Li abundances (^7Li and ^6Li) obtained are $N(\text{Li})/N(\text{H}) = 3.2 \times 10^{-10}$ and 3.3×10^{-10} (with an uncertainty of $\sim 10\%$) for HD82943 and HD91889, respectively. The presence of ^6Li in HD91888 can be rejected with a high level of confidence as the best parabolic fit to the χ^2 as a function of $f(^6\text{Li}) = ^6\text{Li}/^7\text{Li}$ provided $f(^6\text{Li}) = -0.002 \pm 0.006$. On the other hand, the same analysis for

HD82943 yields $f(^6\text{Li}) = 0.126 \pm 0.014$ (Fig. 2). Computations show that this value is not significantly affected by the blue absorption feature owing to the Fe and CN lines. Including a possible unidentified blend¹⁰ at 6,708.02 Å in our synthetic spectra, instead of ^6Li , we will obviously achieve a local fit to the profile. However, the global fit still remains poor because this does not improve the fit at other wavelengths. Synthetic spectra and high-quality spectroscopic observations of similar stars without a Li feature did not reveal any blends from other elements that could affect our conclusions regarding the presence of ^6Li . Additional tests were carried out allowing for large variations of $v \sin i$ and V_{mac} significantly exceeding 3σ limits derived from the Fe lines. We found that these effects do not change our final value of $f(^6\text{Li})$ in HD82943 by more than 0.015.

Our two stars have similar stellar and line-broadening parameters and atmospheric Li content. However, one of them is slightly metal-poor with $[\text{Fe}/\text{H}] = -0.23$ and has no ^6Li , whereas the other is metal-rich ($[\text{Fe}/\text{H}] = +0.32$) and, contrary to expectation, has a $^6\text{Li}/^7\text{Li}$ ratio slightly above the meteoritic value¹² of 0.08. We can rule out the existence of large surface spots as an explanation of the enhanced Li line asymmetry caused by Doppler-shifted components. In fact, this star has a small $v \sin i$ and is photometrically stable on the level of at most 0.006 mag according to Hipparcos data¹⁴. A spot covering only 0.6% of the stellar surface cannot account for the

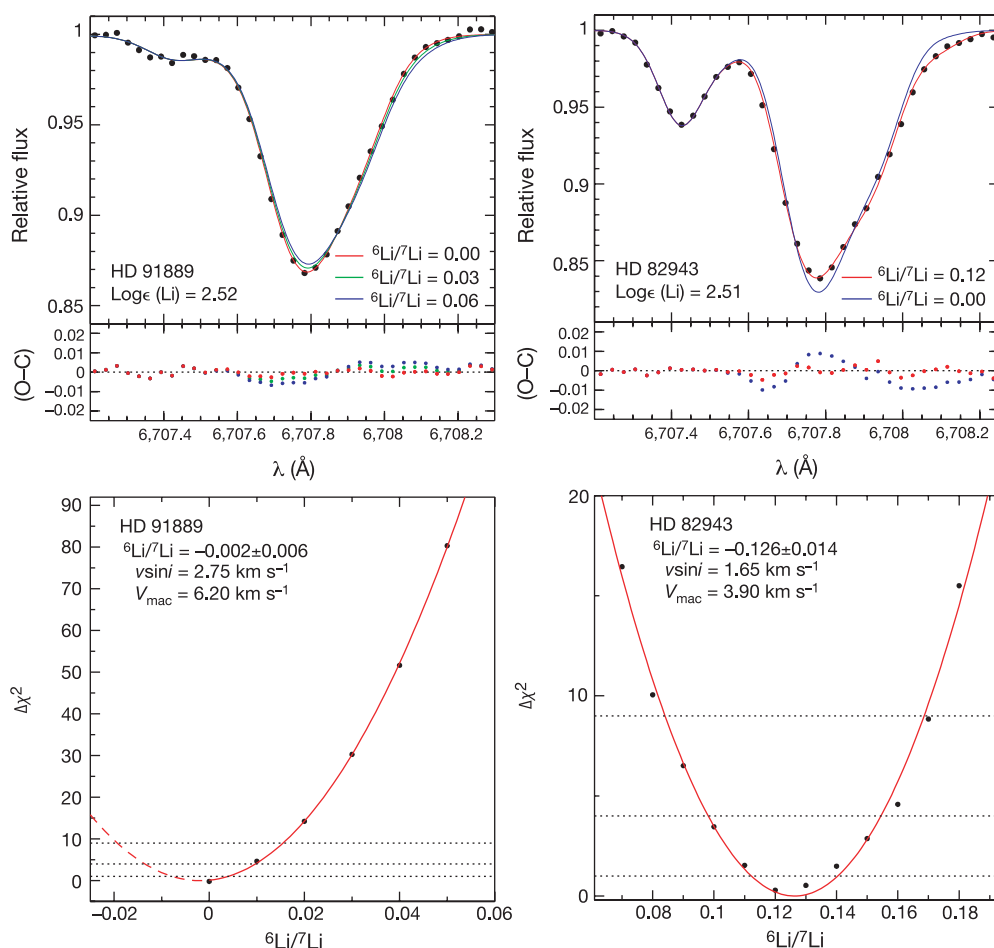


Figure 2 Comparison of the observed (circles) and synthetic (continuous line) spectra corresponding to different $^6\text{Li}/^7\text{Li}$ ratios in our two stars (top) and results from the χ^2 analysis (bottom). The free parameters in the fit are the total Li abundance, the wavelength shift and the ^6Li fraction, defined as $f(^6\text{Li}) = ^6\text{Li}/^7\text{Li}$. The blue wing of the Li line in HD82943 is affected by the absorption feature of the Fe I line at 6,707.433 Å and the CN lines^{17,13} at $\sim 6,707.5$ Å. Therefore, in this case the multi-parameter χ^2 analysis takes into account four additional free parameters: the abundances and the wavelengths of the

Fe and CN absorptions. The absorption feature on the blue wing of the Li line is not so strong in the comparison star and it has a negligible effect on the synthesis. The χ^2 analysis was carried out assuming the mean broadening parameters derived from the Fe lines. The most probable value of $f(^6\text{Li})$ corresponds to the minimum of χ^2 , and $\Delta\chi^2 = 1, 4$ and 9 correspond to the 1, 2 and 3 σ confidence limits (dotted lines). The minimum χ^2 was derived by optimizing all free parameters for each value of $f(^6\text{Li})$ ¹⁰.

observations. In addition, the radial velocities of the star⁶ do not show any external effects that can be produced by surface features. On the other hand, the fact that our target is metal-rich and on the main sequence rules out convective dredge-up as a possible source of the atmospheric Li.

Some observations also make it unlikely that the observed ⁶Li was produced by stellar flares. Particles accelerated by solar flares produce Li isotopes by $\alpha + \alpha$, $\alpha + {}^3\text{He}$ collisions and spallation from C, N, O nuclei¹⁵. An isotopic ratio ${}^6\text{Li}/{}^7\text{Li} = 0.0321 \pm 0.004$ has been advocated for the solar wind from measurements of lunar soil¹⁶. This value exceeds the observational upper limits on the solar photospheric ratio¹⁷, ${}^6\text{Li}/{}^7\text{Li} \leq 0.01$, supporting significant flare production of ⁶Li in the Sun. However, to account for the lunar soil observations most, if not all, of the freshly produced ⁶Li would have to escape from the Sun by the solar wind¹⁵. HD82943 is very similar to the Sun, being only 200 K hotter, 10% more massive and 6 ± 1 Gyr older according to theoretical isochrones¹⁸ and the Hipparcos parallax. Given these characteristics and the low level of chromospheric activity⁶ (similar to that of the Sun), we would expect it to have similar flare activity and ⁶Li production. The current estimates of ⁶Li production in solar flares give values of the order of $\sim 10^{30}$ nuclei for the strongest flares, which take place with a frequency of 30 yr^{-1} and account for most of the ⁶Li production¹⁵. Although it is expected that most of the ⁶Li will escape from the star, the total production of ⁶Li nuclei during 6 Gyr would be $\sim 1.8 \times 10^{41}$, perhaps a factor of 4 higher if we account for the higher C and O abundances in the atmosphere of our target⁷, but still much lower than the 3.2×10^{44} nuclei residing in the $0.01M_{\odot}$ ¹⁹ convection zone of our star. At very early times, during pre-main sequence evolution, solar-type stars have faster rotation, deeper convection zones and can produce much more energetic flares and Li. However, they also have stronger winds aiding the evacuation process, and any residual ⁶Li left in the photosphere will be destroyed by the deep convective mixing.

It has been suggested²⁰ that the ingestion of a planet by an evolved expanding giant star can boost the latter's photospheric Li abundance. Here we discuss the ingestion associated with the fall of planets that have preserved the original level of ⁶Li of the parent protoplanetary and/or protostellar cloud. To explain the presence of

3.2×10^{44} ⁶Li nuclei in the convection zone of HD82943, we estimate that an approximately $2M_J$ giant planet, or a number of planets amounting to this mass, has been engulfed, assuming primordial $N(\text{Li})/N(\text{H}) = 2 \times 10^{-9}$ and $f({}^6\text{Li}) = 0.12$ for the engulfed planets. This can also account for most of the ⁷Li content in our star, meaning that before ingestion the star would have been located in the lower part of the plots in Fig. 3. Our object is currently located in the upper range of the Li abundance distribution. Although this cannot be taken as a proof of planet ingestion, ingestion of a $(1-3)M_J$ giant planet by any star in this diagram (whatever its Li abundance before ingestion) will move its location to the upper envelope. Most stars hosting planets with this temperature domain also have Li abundances in the upper range²¹, but similarly high abundances are found in stars²² that have not been observed to host planets and therefore determination of the ⁶Li content is required to prove the engulfment scenario for each case. The addition of a giant planet with a chemical composition corresponding to the protoplanetary cloud cannot account for the metal-rich nature of our target unless it had a rocky core (a several Earth-mass core consisting mostly of metals). Alternatively, the ingestion of a $3M_{\oplus}$ (where $M_{\oplus}$ is the mass of the Earth) terrestrial planet could provide an explanation for the ⁶Li abundance and contribute to the metallicity enhancement if its iron content were similar to that of the Earth (35.4% by mass²³) and the lithium-to-iron ratio were meteoritic¹². Such a planet would contribute about 17% of the enhancement in Fe nuclei with respect to solar, but any further speculation on the mass of the ingested planet would require knowledge of the initial stellar composition.

The initial ⁶Li/⁷Li ratio of the engulfed planet(s), if any, would be at least equal to the observed value of 0.12 in the atmosphere of the star and might have been preserved during the star's main-sequence evolution if the temperature at the base of the convection zone did not significantly exceed 2×10^6 K. Current models⁵ predict such temperatures to be in the range $(1.5-2.9) \times 10^6$ K for a star with mass $(1.1 \pm 0.1)M_{\odot}$ in the metallicity range $[\text{Fe}/\text{H}] = 0.0-0.3$. Thus, given the errors in stellar parameters and many uncertainties involved in evolutionary models, it is possible that the ⁶Li from the ingested planet(s) was preserved or at least not largely destroyed in the atmosphere of HD82943. Our observations can be used to set

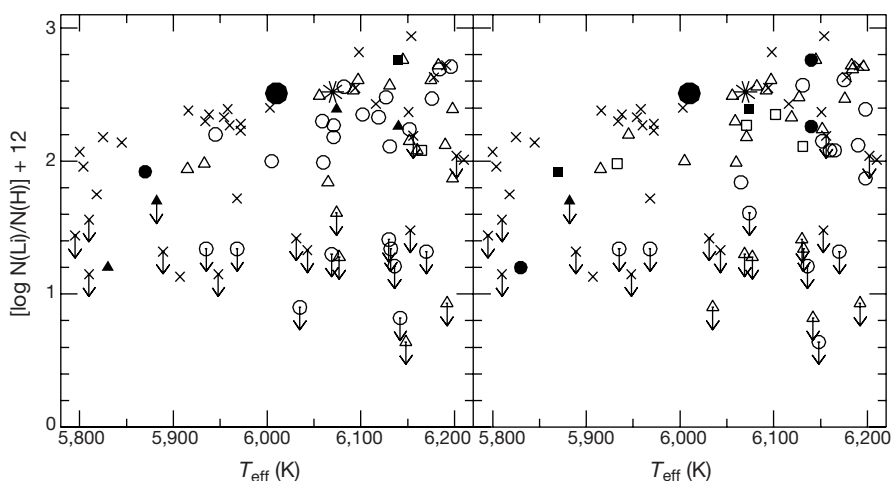


Figure 3 The total lithium abundance of HD82943 (large filled circle) compared with unevolved stars of similar effective temperature (similar mass) for different ages and metallicities^{22,29}. Crosses, stars from the M67 cluster³⁰ (4.5 Gyr old; $[\text{Fe}/\text{H}] \approx 0$). To make a meaningful comparison, only stars with effective temperatures within 200 K of that of HD82943 have been plotted. These stars should have similar superficial convective zones and have suffered similar mixing or diffusion processes. Marked changes in the main lithium depletion mechanisms are expected at temperatures outside this range. Stars have been grouped according to their metallicities (left, $[\text{Fe}/\text{H}] \leq -0.2$ (open

circles), $-0.2 \leq [\text{Fe}/\text{H}] < 0.2$ (open triangles) and $[\text{Fe}/\text{H}] \geq 0.2$ (open squares)) and ages (right, ≤ 4 Gyr, 4–8 Gyr and ≥ 8 Gyr are open circles, triangles and squares, respectively). Small filled symbols are stars hosting planets²¹ and the asterisk denotes our template star. Ages were taken from the original lithium references or inferred using the Geneva isochrones¹⁸. No clear correlation is seen between lithium, effective temperature, age and metallicity. The origin for the large scatter is still unknown and mostly attributed to a variety of lithium depletion mechanisms²¹. Planet(s) accretion is an additional mechanism that can help to understand the observed scatter.

strong constraints on different mixing processes (such as gravitational settling or meridional circulation).

Cosmic-ray nucleosynthesis models²⁴ and observations of interstellar clouds²⁵ indicate that the primordial ratio, $f(^6\text{Li})$, in the protoplanetary matter could have been as high as 0.3–0.5, which in turn would allow either the possibility of some ^6Li depletion in the star or the ingestion of planets with even smaller mass. In this scenario we have assumed that the planet(s) was (were) ingested after the stellar convective envelope had shrunk, 10–30 Myr after stellar birth^{19,5}, and had reached its current mass and depth. The timescale is comparable to the lifetime of protoplanetary disks²⁶ (10–20 Myr) and planetary migration. Our results favour a scenario in which a planet was engulfed owing to the multi-body interactions in the system that can take place over a timescale of 100 Myr²⁷. This is strongly supported by the highly eccentric orbit of the planetary companion of HD82943 (about 0.6). However, we hesitate to generalize it to explain the metal-rich nature^{28,7} of other stars hosting planets. Further Li isotopic ratio studies of a large sample of planet-bearing stars can help to put additional constraints on this scenario.

Note added in proof: A second planet with a minimum mass of 0.88 M_J has been discovered orbiting HD82943, with a period of 220 days (M. Mayor, personal communication). This detection has been done with the CORALIE spectrograph at the ESO La Silla Observatory. □

Received 3 November 2000; accepted 21 March 2001.

- Weidenschilling, S. J. & Marzari, F. Gravitational scattering as a possible origin for giant planets at small distances. *Nature* **384**, 619–621 (1996).
- Lin, D. N., Bodenheimer, P. & Richardson, D. C. Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature* **380**, 606–607 (1996).
- Forestini, M. Low mass stars: Pre-main sequence evolution and nucleosynthesis. *Astron. Astrophys.* **285**, 473–488 (1994).
- Andersen, J., Gustafsson, B. & Lambert, D. L. The lithium isotope ratio in F and G stars. *Astron. Astrophys.* **136**, 65–73 (1984).
- Rebolo, R., Crivellari, L., Castell, F., Foing, B. & Beckman, J. E. Lithium abundances and $^7\text{Li}/^6\text{Li}$ ratios in late-type population I field dwarfs. *Astron. Astrophys.* **166**, 195–203 (1986).
- Naef, D. *et al.* The CORALIE survey for southern extrasolar planets V. 4 new extrasolar planets. *Astron. Astrophys.* (in the press).
- Santos, N. C., Israelian, G. & Mayor, M. Chemical analysis of 8 recently discovered extra-solar planet host stars. *Astron. Astrophys.* **363**, 228–238 (2000).
- Smith, V. V., Lambert, D. & Nissen, P. E. Isotopic lithium abundances in nine halo stars. *Astrophys. J.* **506**, 405–423 (1998).
- Hobbs, L. M., Thorburn, J. A. & Rebull, L. M. Lithium isotope ratios in halo stars. III. *Astrophys. J.* **523**, 797–804 (1999).
- Nissen, P. E., Lambert, D., Primas, F. & Smith, V. V. Isotopic lithium abundances in five metal-poor disk stars. *Astron. Astrophys.* **348**, 211–221 (1999).
- Kurucz, R. L., Furenlid, I., Brault, J. & Testerman, L. Solar flux atlas from 296 to 1300 nm. *NOAO Atlas 1* (Harvard Univ. Press, Cambridge, Massachusetts, 1984).
- Anders, E. & Grevesse, N. Abundances of the elements—meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
- Lambert, D., Smith, V. V. & Heath, J. Lithium in the barium stars. *Publ. Astron. Soc. Pacif.* **105**, 568–573 (1989).
- Perryman, M. A. C. *et al.* The Hipparcos catalogue. *Astron. Astrophys. Lett.* **323**, 49–52 (1997).
- Ramaty, R., Tatischeff, V., Thibaud, J. P., Kozlovsky, B. & Mandzhavidze, N. ^6Li from solar flares. *Astrophys. J. Lett.* **534**, 207–210 (2000).
- Chaussidon, M. & Robert, F. Lithium nucleosynthesis in the sun inferred from the solar-wind $^7\text{Li}/^6\text{Li}$ ratio. *Nature* **402**, 270–273 (1999).
- Müller, E. A., Peytremann, E. & De La Reza, R. The solar lithium abundance. II. *Solar Phys.* **41**, 53–65 (1975).
- Schaefer, D., Charbonnel, C., Meynet, G., Maeder, A. & Schaller, G. Grids of stellar models—Part four—from 0.8 $M_\odot$ to 120 $M_\odot$ at $Z = 0.040$. *Astron. Astrophys. Suppl.* **102**, 339–342 (1993).
- D’Antona, F. & Mazzitelli, I. New pre-main sequence tracks for $M \leq 2.5 M_\odot$ as tests of opacities and convection models. *Astrophys. J. Suppl.* **90**, 467–500 (1994).
- Alexander, J. B. A possible source of lithium in the atmospheres of some red giants. *Observatory* **87**, 238–240 (1967).
- Ryan, S. G. The host stars of extrasolar planets have normal lithium abundances. *Mon. Not. R. Astron. Soc.* **316**, L35–L39 (2000).
- Balachandran, S. Lithium depletion and rotation in main-sequence stars. *Astrophys. J.* **354**, 310–332 (1990).
- Glass, B. P. *Introduction To Planetary Geology* (Cambridge Univ. Press, Cambridge, 1982).
- Fields, B. D. & Olive, K. A. The evolution of ^6Li in standard cosmic ray nucleosynthesis. *New Astron.* **4**, 255–263 (1999).
- Knauth, D. C., Federman, S. R., Lambert, D. L. & Crane, P. Newly synthesized lithium in the interstellar medium. *Nature* **405**, 656–658 (2000).
- Thi, W. F. *et al.* Substantial reservoirs of molecular hydrogen in the debris disks around young stars. *Nature* **409**, 60–63 (2001).

- Levison, H. F., Lissauer, J. J. & Duncan, M. J. Modeling the diversity of outer planetary systems. *Astron. J.* **373**, 1998–2014 (1998).
- Gonzalez, G., Laws, C., Tyagi, S. & Reddy, B. E. Parent stars of extrasolar planets VI: Abundance analyses of 20 new systems. *Astron. J.* **121**, 432–452 (2001).
- Boesgaard, A. M. & Tripicco, M. Lithium in early F dwarfs. *Astrophys. J.* **303**, 724–739 (1986).
- Jones, B. F., Fisher, D. & Soderblom, D. R. The evolution of the lithium abundances of solar-type stars. VIII. M67 (NGC 2682). *Astron. J.* **117**, 330–338 (1999).

Acknowledgements

These observations were made possible through the DDT time granted on the VLT Kueyen by ESO. We thank the Swiss National Science Foundation and the Spanish Ministry of Science and Technology for continuous support for this project. Support from the Fundação para a Ciência e Tecnologia, Portugal, and to N.C.S. in the form of a scholarship is acknowledged.

Correspondence and requests for materials should be addressed to G.I. (e-mail: gil@lliac.es).

A complementarity experiment with an interferometer at the quantum–classical boundary

P. Bertet, S. Osnaghi, A. Rauschenbeutel, G. Nogues, A. Auffeves, M. Brune, J. M. Raimond & S. Haroche

Laboratoire Kastler Brossel, Département de Physique, Ecole Normale Supérieure, 24 rue Lhomond, F-75231, Paris Cedex 05, France

To illustrate the quantum mechanical principle of complementarity, Bohr¹ described an interferometer with a microscopic slit that records the particle’s path. Recoil of the quantum slit causes it to become entangled with the particle, resulting in a kind of Einstein–Podolsky–Rosen pair². As the motion of the slit can be observed, the ambiguity of the particle’s trajectory is lifted, suppressing interference effects. In contrast, the state of a sufficiently massive slit does not depend on the particle’s path; hence, interference fringes are visible. Although many experiments illustrating various aspects of complementarity have been proposed^{3–9} and realized^{10–18}, none has addressed the quantum–classical limit in the design of the interferometer. Here we report an experimental investigation of complementarity using an interferometer in which the properties of one of the beam-splitting elements can be tuned continuously from being effectively microscopic to macroscopic. Following a recent proposal¹⁹, we use an atomic double-pulse Ramsey interferometer²⁰, in which microwave pulses act as beam-splitters for the quantum states of the atoms. One of the pulses is a coherent field stored in a cavity, comprising a small, adjustable mean photon number. The visibility of the interference fringes in the final atomic state probability increases with this photon number, illustrating the quantum to classical transition.

Interferences in ordinary space are easier to understand than those of the abstract space of quantum states, so it is illuminating to make the analogy between the Ramsey design²⁰ and a standard optical interferometer. Instead of Bohr’s original design¹ (Fig. 1), we describe here a Mach–Zehnder interferometer³ which provides a closer analogy with our Ramsey experiment (Fig. 2a). A photon beam is split into two paths *a* and *b* and recombined by two beam splitters *B*₁ and *B*₂. The quantum amplitudes associated with these paths present a phase difference ϕ , swept by a retarding element. If *B*₁ and *B*₂ are macroscopic, the probability for detecting the particle in detector *D* exhibits a sinusoidal modulation as a function of ϕ . Suppose now that *B*₂ is a massive classical object, and *B*₁ is a light plate rotating around an axis perpendicular to the interferometer

plane. When the particle interacts with B_1 , it is, with a 50% probability, either transmitted along path a (the plate does not move) or reflected into path b (the plate receives a momentum kick, resulting in a coherent state of motion). The particle + B_1 system evolves into the combined state:

$$|\Psi\rangle = \left(1/\sqrt{2}\right)(|\Psi_a\rangle + |\Psi_{B_1}(a)\rangle + |\Psi_b\rangle + |\Psi_{B_1}(b)\rangle) \quad (1)$$

where $|\Psi_a\rangle$ and $|\Psi_b\rangle$ represent the particle's wave packets in paths a and b and $|\Psi_{B_1}(a)\rangle$ and $|\Psi_{B_1}(b)\rangle$ the corresponding final states of B_1 . If B_1 is light enough, it stores unambiguous information about the particle's path: $\langle\Psi_{B_1}(b)|\Psi_{B_1}(a)\rangle = 0$, and $|\Psi\rangle$ is maximally entangled. If B_1 is a heavy macroscopic object insensitive to the particle's momentum kick, $|\Psi_{B_1}(a)\rangle = |\Psi_{B_1}(b)\rangle$ and $|\Psi\rangle$ is an unentangled product state. The probability for detecting the particle in D is:

$$P(\phi) = \frac{1}{2}(1 + \text{Re}(\langle\Psi_{B_1}(b)|\Psi_{B_1}(a)\rangle \exp(i\phi))) \quad (2)$$

The fringe contrast is the modulus of the scalar product of the two beam-splitter final states and directly measures the degree of entanglement between the particle and B_1 .

Our Ramsey interferometer, sketched in Fig. 2b, presents an analogy with the Mach–Zehnder device. We describe here only the main features of the set-up; more experimental details can be found elsewhere^{21,22}. Rubidium atoms prepared in the circular Rydberg state of principal quantum number 51 (level e) are sent one by one through the interferometer with a velocity of 500 m s^{-1} . Their position is known within 1 mm at all times during their transit across the apparatus. The atoms undergo a transition between e and the lower-energy circular Rydberg level g (principal quantum number 50), at frequency $\nu_{eg} = 51.1 \text{ GHz}$. This transition is induced by two coherent pulses R_1 and R_2 mixing with equal weights e and g ($\pi/2$ pulses), at two different times separated by a delay $T = 24 \mu\text{s}$. The final atomic energy state is detected in D. The transition probability P_g is reconstructed by averaging a large number of single-atom events.

The pulse R_1 is produced by a small coherent field of complex amplitude α stored in a superconducting Fabry–Perot resonator C (photon damping time $T_{\text{cavity}} = 1 \text{ ms}$). This field is created, before the atom enters the interferometer, by a pulsed coherent microwave

source S. The field mode, with a gaussian profile (6 mm waist), is crossed by the atom along a direction perpendicular to the cavity axis. The coherent field state, $|\alpha\rangle = \sum_n C_n |n\rangle$, is a superposition of photon numbers with a Poisson distribution. The probability amplitude for having n photons is $C_n = \exp(-|\alpha|^2/2)\alpha^n/\sqrt{n!}$. The mean photon number is $N = |\alpha|^2$, with a variance $\Delta N = \sqrt{N} = |\alpha|$ and a phase fluctuation³ $\Delta\Phi = 1/|\alpha|$, the minimum value allowed by the Heisenberg uncertainty relation $\Delta N \Delta\Phi \geq 1$.

We determine N by the measurement of the light shift experienced by an auxiliary atom which is crossing the cavity while being non-resonant with its mode²³. By varying the source amplitude, we tune N from 0 to large values. The $\pi/2$ pulse condition is satisfied by adjusting the duration of the atom–field interaction for each N value. A small electric field is applied between the mirrors of the cavity while the atom is crossing it. In this way, we tune by Stark effect the e – g transition either in or out of resonance with the cavity mode. The atom enters the cavity out of resonance. No evolution occurs until the resonance condition is realised by Stark tuning, at a time such that the $\pi/2$ pulse is achieved when the atom leaves the cavity. The duration of this pulse is a decreasing function of N . The large value of the vacuum Rabi frequency at the cavity centre²⁴, $\Omega/2\pi = 49 \text{ kHz}$, ensures that the $\pi/2$ pulse can be achieved even when $N = 0$. The possibility of inducing a $\pi/2$ pulse with the vacuum field, in a time much shorter than T_{cavity} , is an essential feature of cavity experiments with circular Rydberg atoms²⁵.

After R_1 , the combined atom + beam splitter state can be written, in a form similar to equation (1), as:

$$|\Psi\rangle = \left(1/\sqrt{2}\right)(|e\rangle + |\alpha_e\rangle + |g\rangle + |\alpha_g\rangle) \quad (3)$$

with

$$\begin{aligned} |\alpha_e\rangle &= \sqrt{2} \left(\sum_n C_n \cos(\Omega\sqrt{n+1}t_\alpha) |n\rangle \right) \\ |\alpha_g\rangle &= \sqrt{2} \left(\sum_n C_n \sin(\Omega\sqrt{n+1}t_\alpha) |n+1\rangle \right) \end{aligned} \quad (4)$$

where t_α is an effective atom–cavity interaction time defined by:

$$\sum_n |C_n|^2 \cos^2(\Omega\sqrt{n+1}t_\alpha) = 1/2 \quad (5)$$

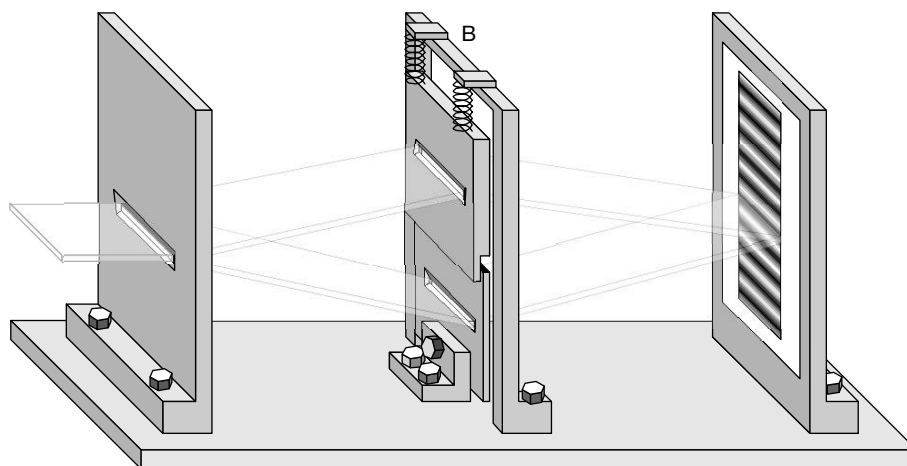


Figure 1 Sketch of an interferometer with a recoiling slit, adapted from Bohr¹. Particles collimated by the fixed slit at left cross a double slit assembly B and are collected on a detecting plate at right. The bottom slit in B, firmly bolted on the ground plate, cannot move. The upper slit is suspended by springs and assumed to be initially in the ground state of its oscillator motion. When its mass is small enough, the momentum kick imparted by the deflected particle is larger than the initial slit's momentum fluctuations. The state of the slit is entangled with the one of the particle. This provides which-path information and

suppresses the interference fringes. When the upper slit is a macroscopic heavy object, its state is not altered by the particle and fringes are observed. By increasing the mass of the slit while keeping the spring stiffness fixed, one could, in principle, continuously span the transition between the two limiting cases discussed by Bohr of a recoiling or fixed slit. The drawing is typical of the semi-serious style adopted by Bohr in his description of thought experiments (in Bohr's original drawings, the various elements were represented separately).

Equation (4) shows that each n -photon component of the coherent state induces a Rabi oscillation between e and g at frequency $\Omega\sqrt{n+1}$ ^{24,25}. The global system's evolution results from the superposition of these Rabi oscillations. In the quantum limit ($N = 0$), $|\alpha_e\rangle = |0\rangle$ and $|\alpha_g\rangle = |1\rangle$ are orthogonal states. In the classical limit, $N \gg 1$, ΔN becomes negligible compared to N . According to equation (5), the $\cos(\Omega\sqrt{n+1}t_\alpha)$ and $\sin(\Omega\sqrt{n+1}t_\alpha)$ functions are approximately $1/\sqrt{2}$, and from equation (4), $|\alpha_e\rangle \approx |\alpha_g\rangle \approx |\alpha\rangle$. The field is thus not significantly altered by the atom. The quantum to classical evolution of the atom–field entanglement described by equation (3) thus exhibits all the qualitative features discussed for the Mach–Zehnder interferometer.

The pulse R_2 is generated by S and fed, via an auxiliary wave guide, into a standing wave independent of C. This standing wave, determined by the structure of the apparatus metallic boundaries, presents an antinode at a distance of 18 mm from the cavity axis. When the atom crosses this antinode, S is switched on for a short

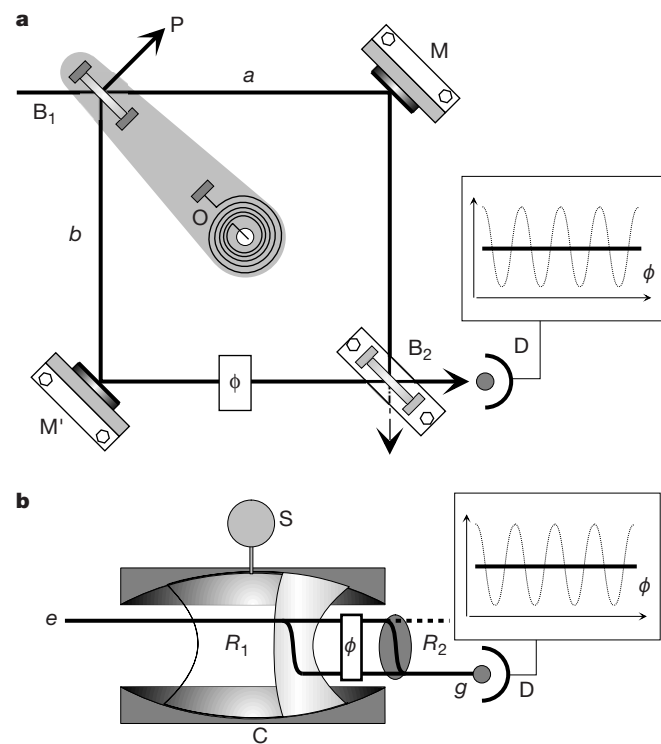


Figure 2 Mach–Zehnder and Ramsey versions of Bohr's experiment. **a**, In the Mach–Zehnder, the particle trajectory is separated by beam splitter B_1 into paths a and b , folded by mirrors M and M' and recombined by beam splitter B_2 into detector D . The other output port (dashed arrow) is not used. A dephasing element tunes the relative phase ϕ between the paths. B_1 rotates around an axis perpendicular to the interferometer plane, crossing it at the centre O of the B_1MB_2M' square. A spring provides a restoring force. The moving assembly is initially in its ground state of motion. If B_1 has a large mass, fringes are visible (dotted lines in inset). If B_1 is microscopic, its recoil records the reflection of the particle into path b , washing out the fringes (solid lines). **b**, Ramsey set-up. A Rydberg atom's state is split by microwave pulse R_1 into two energy levels e and g , then recombined by pulse R_2 downstream, before being detected by field-ionization in D . We use a mixed representation combining real space with the space of the atomic energy states. The beam-splitting effect, a spatial separation in the Mach–Zehnder, is now an internal mixture of states. Interferences are obtained in the probability for finding the atom in g . A field pulse between R_1 and R_2 tunes the relative phase ϕ of the interfering amplitudes (Stark effect). If the R_1 field, stored in a long-damping-time cavity C , is macroscopic, Ramsey fringes are visible (dotted line in the inset). If the R_1 field is microscopic, its photon number records information about the atom's path, suppressing the interference (solid line). The Gaussian mode of the cavity field is represented schematically; the part where the atom is resonant is shaded. The cavity mode waist and the R_1 – R_2 distance are exaggerated.

time ($1\ \mu\text{s}$), realising the $\pi/2$ pulse condition. The damping time of this standing wave is very short (less than 10 ns). The photons of the R_2 pulse are regenerated many times during the interaction of the atom with this field. Hence, the R_2 field does not get entangled with the atomic state. It can be considered as a classical beam splitter²⁶ performing the single-particle transformations $|e\rangle \Rightarrow (1/\sqrt{2})(|e\rangle + |g\rangle)$ and $|g\rangle \Rightarrow (1/\sqrt{2})(|g\rangle - |e\rangle)$.

Between R_1 and R_2 , the probability amplitudes for finding the atom in e or g accumulate a quantum phase difference ϕ . In order to tune it, we subject the atom to a 2- μs pulse of electric field applied across the cavity mirrors, which shifts the atomic transition frequency and thus ϕ by a variable amount ($0 < \phi < 3\pi$). Taking into account the action of the beam splitters and this phase accumulation, the transition probability between levels e and g may be written as:

$$P_g(\phi) = \frac{1}{2}(1 + \text{Re}(\langle \alpha_e | \alpha_g \rangle \exp(i\phi))) \quad (6)$$

In Fig. 3a we plot the fringe signals obtained for different amplitudes of the R_1 field, varying from the vacuum ($N = 0$) to $N = 12.8$ ($|\alpha| = 3.5$). In the vacuum case, the fringes are completely washed out: the cavity mode, which undergoes the $0 \Rightarrow 1$ photon transition when the atom flips from e to g , stores 'which path' information. When N increases, the fringes appear, their contrast reaching 72% for $N = 12.8$. This reflects the progressive loss of which-path information, as the beam splitter becomes more and more classical. The maximum fringe contrast is smaller than the ideal 100% given by equation (6), by a factor η due to various experimental imperfections^{21,22}. Figure 3b shows the variations of the fringe contrast as a function of N . The points are experimental and the curve represents the theoretical variations of $\eta |\langle \alpha_e | \alpha_g \rangle|$ where $\eta = 0.75$ is determined by adjusting the theoretical curve on

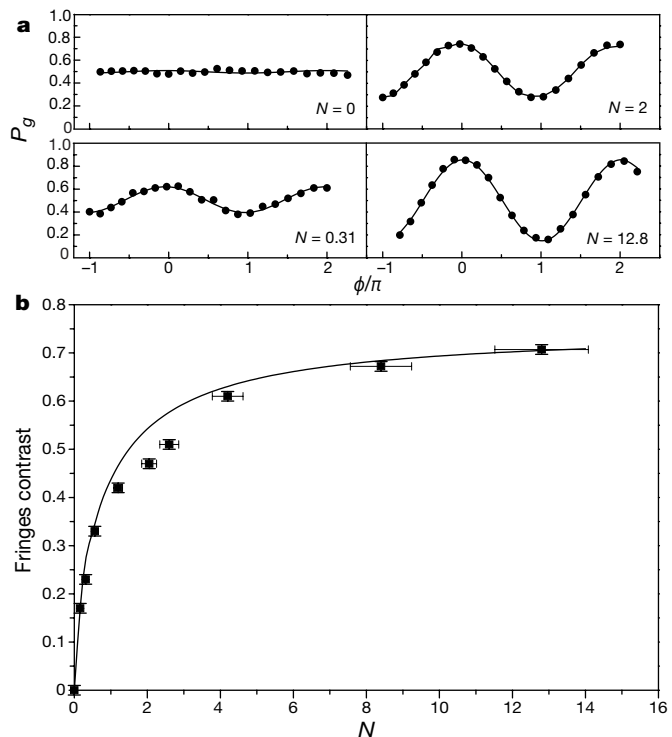


Figure 3 From quantum to classical interferometer. **a**, Ramsey interference signal recorded for various mean photon numbers N in the R_1 pulse. The progressive evolution from the quantum to the classical beam-splitter case is clearly observed. **b**, Fringe contrast as a function of the mean photon number N in R_1 . The points are experimental. The line represents the theoretical variation of the modulus of the beam-splitter final-states scalar product, multiplied by a fixed factor η that accounts for interferometer imperfections.

the experimental point at $N = 12.8$. There is excellent agreement between experiment and theory.

Here, as already proposed^{4,19}, the which-path information is stored in the photon number. It could also be coded in the phase of a coherent field, as suggested^{15,27} and demonstrated²⁸. In ref. 28, the field interacted in a non-resonant way with Rydberg atoms, between two classical pulses. The atomic index of refraction, which is different for e and g , left the which-path information encoded, not in the photon number, but in the field phase. Contrary to what we observe now (see above), fringes then disappeared for large field amplitudes. The field was then a measuring apparatus, distinct from the interferometer, recording information about the particle's path (that is, its energy) only when it was a classical macroscopic object. Here, the cavity field is part of the interferometer, an apparatus aimed at measuring atomic coherences. The field is a 'well behaved' beam splitter, producing highly contrasted fringes when it contains many photons. Both experiments are in agreement with Bohr's definition of a genuine measuring device, whose parts must be classical.

We can also analyse our experiment in terms of a phase-photon number uncertainty relation. The fringe contrast measures the phase correlation between the two beam splitters, as R_1 imprints its phase on an atomic state superposition, read out later by R_2 . The R_2 field has a well defined phase, whereas R_1 presents quantum phase fluctuations $\Delta\Phi$, of the order of $1/\Delta N$. In order to get which-path information, we need $\Delta N < 1$, making a one-photon change in R_1 detectable. This entails $\Delta\Phi > 1$, resulting in the washing-out of phase correlations between R_1 and R_2 . The argument is only qualitative here, as the phase is not a real quantum mechanical operator³ (similar Heisenberg uncertainty arguments, based on the $\Delta x \Delta p > 1$ relation, can be invoked to analyse the Mach-Zehnder thought experiment). In this way, we can easily explain the fringe contrast for $N = 0$ and N large. For intermediate N values, the fundamental atom- R_1 entanglement approach is more precise. Interpreted in a superficial way, the Heisenberg uncertainty explanation seems to describe the R_1 -atom interaction as an irreversible noisy perturbation blurring away quantum interferences. This simplistic view is deceptive, however, and the coding of which-path information can be, as we now show, a fully reversible process.

We performed an experiment involving two successive interactions with the same quantum beam splitter, in a configuration such that the which-path information stored during the first interaction was erased during the second, leading to an unconditional restoration of the fringes. The Mach-Zehnder analogue of this experiment is sketched in Fig. 4a and the actual set-up in Fig. 4b. We now assume that the two beam splitters are microscopic quantum plates rigidly coupled together by a rotating assembly moving around O. The B_1 and B_2 plates are kicked when the particle follows the paths b and a , respectively, resulting in both cases in the same final rotation of the assembly (arrows in Fig. 4a). The which-path information acquired by B_1 is 'erased' when the particle reaches B_2 , a situation reminiscent of the 'quantum eraser'^{23,17,19,29,30}. The fringes should then be visible. In usual quantum-eraser procedures, however, the which-path recording system is actively manipulated, then measured. Conditional interferences appear, whose phase depends upon the outcome of the which-path detector measurement. If this measurement is not performed, fringes do not show up. Here, on the contrary, the information is unconditionally erased by the second interaction with the beam splitter.

The Ramsey version of this experiment (Fig. 4b) involves two successive pulses induced by the quantum field in C. The cavity is initially in vacuum state. The first pulse occurs when the atom enters the mode waist and the second just before it leaves the waist, a time $T = 16 \mu\text{s}$ later. The interaction with the quantum field is interrupted in between by Stark switching. An additional 2- μs pulse of electric field is used, as in the previous experiment, to sweep the phase difference ϕ between the two paths. The first Ramsey pulse

transforms the atom + field state into the entangled superposition described by equation (3) with $|\alpha_e\rangle = |0\rangle$ and $|\alpha_g\rangle = |1\rangle$. The interferometer then stores which-path information. Just before the second pulse, the atom + field system is in the combined state:

$$|\Psi'\rangle = \left(1/\sqrt{2}\right)(\exp(i\phi)|e\rangle|0\rangle + |g\rangle|1\rangle) \quad (7)$$

The second quantum pulse mixes again coherently $|e,0\rangle$ and $|g,1\rangle$ transforming this superposition into:

$$|\Psi''\rangle = \frac{1}{2}[|g\rangle|1\rangle(1 + \exp(i\phi)) - |e\rangle|0\rangle(1 - \exp(i\phi))] \quad (8)$$

The system's state contains two terms corresponding to the atom ending up in g , with a phase difference ϕ between them. In both terms, the field is in the same one-photon state. The likeness to the Mach-Zehnder case is noticeable: the photon in the cavity replaces the momentum kick stored in the beam-splitter assembly. The final

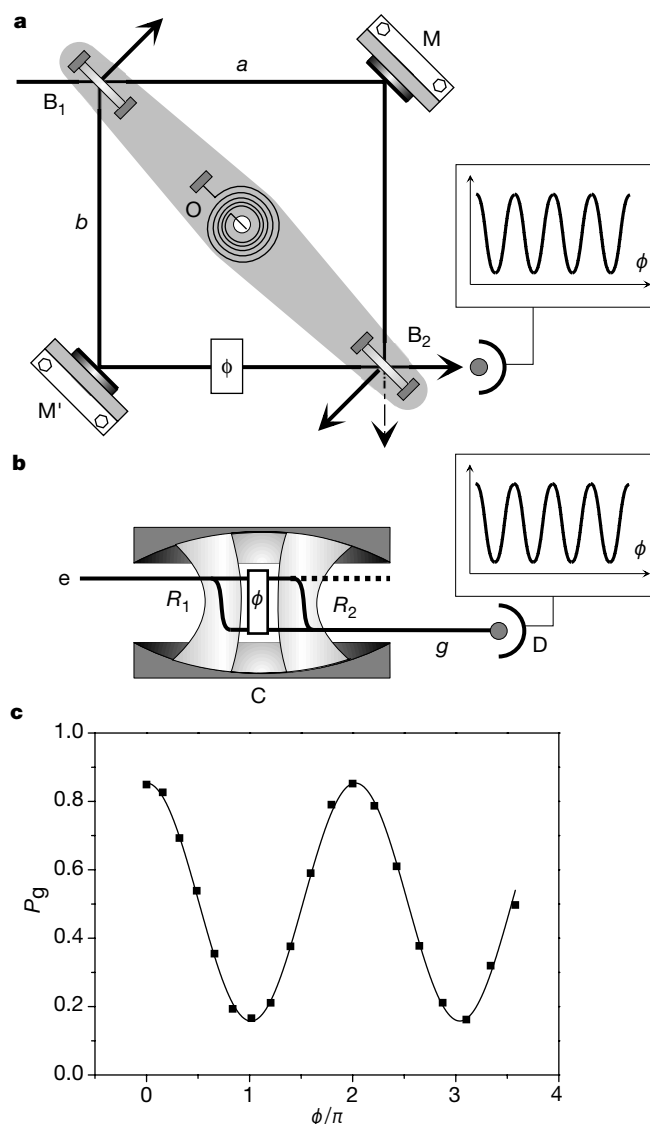


Figure 4 The unconditional quantum-eraser experiment. **a**, Mach-Zehnder's version: B_1 and B_2 are rigidly coupled to the same assembly rotating around O. Both paths result in the same angular momentum kick of the B_1 - B_2 system. Fringes should appear. **b**, Actual Ramsey experiment. R_1 and R_2 are produced by the same quantum mode in C, initially empty. The coupling of the atom is interrupted between R_1 and R_2 by Stark switching. In both paths ending in level g , the atom finally releases one photon in C: the which-path information, produced by the first pulse, is 'erased' by the second. **c**, Interference signal obtained under these conditions exhibiting a large fringe contrast.

information contained in the field does not distinguish the two paths leading to g . The two corresponding amplitudes interfere. This results in large contrast fringes on P_g , as shown in Fig. 4c. We note that these fringes are visible despite the overall random phase of the pulses. The fringe contrast is sensitive to the phase correlation of the Ramsey fields between the times 0 and T , which is non-vanishing as long as $T < T_{\text{cavity}}$.

We have shown that a coherent field stored in a low-loss cavity realises a versatile beam splitter for atomic interferometry experiments. It can evolve continuously from a microscopic device, able to store which-path information, into a macroscopic system with a large number of quanta, insensitive to the interaction with the interfering particle. When the beam splitter is quantum, its coupling with the particle results in maximum entanglement between the two systems. When the same quantum beam-splitter is used twice in the interferometer, the quantum information is erased and the fringes restored. By allowing us to explore the transition from microscopic to macroscopic measuring devices, these experiments shed light on the quantum-classical boundary. □

Received 22 December 2000; accepted 7 March 2001.

- Bohr, N. in *Albert Einstein: Philosopher Scientist* (ed. Schilpp, P. A.) 200–241 (Library of Living Philosophers, Evanston, 1949); reprinted in *Quantum Theory and Measurement* (eds Wheeler, J. A. & Zurek, W. H.) 9–49 (Princeton Univ. Press, Princeton, 1983).
- Einstein, A., Podolsky, B. & Rosen, N. Can quantum mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
- Scully, M. O. & Zubairy, M. S. *Quantum Optics* (Cambridge Univ. Press, Cambridge, UK, 1997).
- Scully, M. O., Englert, B. G. & Walther, H. Quantum optical tests of complementarity. *Nature* **351**, 111–116 (1991).
- Haroche, S., Brune, M. & Raimond, J. M. Manipulations of optical fields by atomic interferometry: quantum variations on a theme by Young. *Appl. Phys. B* **54**, 355–365 (1992).
- Englert, B. G. Fringe visibility and which-way information: an inequality. *Phys. Rev. Lett.* **77**, 2154–2157 (1996).
- Englert, B. G., Scully, M. O. & Walther, H. Complementarity and uncertainty. *Nature* **375**, 367 (1995).
- Storey, P., Tan, S., Collet, M. & Walls, D. Complementarity and uncertainty. *Nature* **375**, 368 (1995).
- Wiseman, H. *et al.* Non local momentum transfer in welcher Weg measurements. *Phys. Rev. A* **56**, 55–75 (1997).
- Badurek, G., Rausch, H. & Tuppinger, D. Neutron interferometric double resonance experiments. *Phys. Rev. A* **34**, 2600–2608 (1985).
- Grangier, P., Roger, G. & Aspect, A. Experimental evidence for photon anticorrelation effect on a beam splitter: a new light on single-photon interference. *Europhys. Lett.* **1**, 173–179 (1986).
- Pfau, T. *et al.* Loss of spatial coherence by a single spontaneous emission. *Phys. Rev. Lett.* **73**, 1223–1227 (1994).
- Chapman, M. S. *et al.* Photon scattering from atoms in an atom interferometer: coherence lost and regained. *Phys. Rev. Lett.* **75**, 3783–3787 (1995).
- Eichman, U. *et al.* Young's interference experiment with light scattered from two atoms. *Phys. Rev. Lett.* **70**, 2359–2362 (1993).
- Dürr, S., Nonn, T. & Rempe, G. Origin of quantum-mechanical complementarity probed by a “which way” experiment in an atom interferometer. *Nature* **395**, 33–37 (1998).
- Dürr, S., Nonn, T. & Rempe, G. Fringe visibility and which-way information in an atom interferometer. *Phys. Rev. Lett.* **81**, 5705–5709 (1998).
- Herzog, T. J., Kwiat, P. G., Weinfurter, H. & Zeilinger, A. Complementarity and the quantum eraser. *Phys. Rev. Lett.* **75**, 3034–3037 (1995).
- Bucks, E., Schuster, R., Heiblum, M., Mahalu, D. & Umansky, V. Dephasing in electron interference by a “which-path” detector. *Nature* **391**, 871–874 (1998).
- Zheng, S. B. A simplified scheme for testing complementarity and realising quantum eraser. *Opt. Comm.* **173**, 265–267 (2000).
- Ramsey, N. *Molecular Beams* (Oxford Univ. Press, Oxford, UK, 1985).
- Nogues, G. *et al.* Seeing a single photon without destroying it. *Nature* **400**, 239–242 (1999).
- Rauschenbeutel, A. *et al.* Coherent operation of a tunable quantum phase gate. *Phys. Rev. Lett.* **83**, 5166–5169 (1999).
- Brune, M. *et al.* From Lamb shift to light shifts: vacuum and sub-photon cavity fields measured by atomic interferometry. *Phys. Rev. Lett.* **72**, 3339–3342 (1994).
- Brune, M. *et al.* Quantum Rabi oscillation: a direct test of field quantization in a cavity. *Phys. Rev. Lett.* **76**, 1800–1803 (1996).
- Haroche, S. & Raimond, J. M. in *Cavity Quantum Electrodynamics* (ed. Berman, P.) 123–170 (Academic, New York, 1992).
- Kim, J. I. *et al.* Classical behaviours with small quantum numbers: the physics of Ramsey interferometry of Rydberg atoms. *Phys. Rev. Lett.* **82**, 4737–4740 (1999).
- Gerry, C. C. Complementarity and quantum erasure with dispersive atom-field interaction. *Phys. Rev. A* **53**, 1179–1182 (1996).
- Brune, M. *et al.* Observing the progressive decoherence of the meter in a quantum measurement. *Phys. Rev. Lett.* **77**, 4887–4890 (1996).
- Scully, M. O. & Söhl, K. Quantum eraser: a proposed photon correlation experiment concerning observation and “delayed choice” in quantum mechanics. *Phys. Rev. A* **25**, 2208–2213 (1982).
- Kwiat, P. G., Steinberg, A. M. & Chiao, R. Y. Observation of a “quantum eraser”: a revival of coherence in a 2-photon interference experiment. *Phys. Rev. A* **45**, 7729–7739 (1992).

Acknowledgements

Laboratoire Kastler Brossel is Unité Mixte de Recherches of Ecole Normale Supérieure, Université P. et M. Curie and Centre National de la Recherche Scientifique. This work was supported by the Commission of the European Community and by the Japan Science and Technology Corporation (International Cooperative Research Project, Quantum Entanglement Project).

Correspondence and requests for materials should be addressed to S.H. (e-mail: haroche@lkb.ens.fr).

Semiconducting non-molecular nitrogen up to 240 GPa and its low-pressure stability

Mikhail I. Erements, Russell J. Hemley, Ho-kwang Mao & Eugene Gregoryanz

Geophysical Laboratory and Center for High Pressure Research, Carnegie Institution of Washington, 5251 Broad Branch Road NW, Washington DC 20015, USA

The triple bond of diatomic nitrogen has among the greatest binding energies of any molecule. At low temperatures and pressures, nitrogen forms a molecular crystal in which these strong bonds co-exist with weak van der Waals interactions between molecules, producing an insulator with a large band gap¹. As the pressure is raised on molecular crystals, intermolecular interactions increase and the molecules eventually dissociate to form monoatomic metallic solids, as was first predicted for hydrogen². Theory predicts that, in a pressure range between 50 and 94 GPa, diatomic nitrogen can be transformed into a non-molecular framework or polymeric structure with potential use as a high-energy-density material^{3–5}. Here we show that the non-molecular phase of nitrogen is semiconducting up to at least 240 GPa, at which pressure the energy gap has decreased to 0.4 eV. At 300 K, this transition from insulating to semiconducting behaviour starts at a pressure of approximately 140 GPa, but shifts to much higher pressure with decreasing temperature. The transition also exhibits remarkably large hysteresis with an equilibrium transition estimated to be near 100 GPa. Moreover, we have succeeded in recovering the non-molecular phase of nitrogen at ambient pressure (at temperatures below 100 K), which could be of importance for practical use.

McMahan and LeSar³ first predicted that nitrogen would transform from the molecular state, in which each atom is triple-bonded with one near neighbour, to a monoatomic structure in which each atom is bonded to three nearest neighbours by single covalent bonds. This unusual transition is also interesting from the point of view of the chemistry of nitrogen and its applications. Only small molecules with two or three nitrogen atoms in a row are known (the only exception being perhaps $N_2^+AsF_6^-$; ref. 6). These may form high-energy-density materials because the transformation from polymerized nitrogen to diatomic molecular nitrogen is accompanied by a very large energy release⁶ (the average bond strength of the N–N single bond is 160 kJ mol^{−1} and that of the triple bond is 954 kJ mol^{−1}). The calculations^{3–5} predicted this transformation to occur at pressures of 50–94 GPa with a large volume change of approximately 35%; this prediction has been examined in additional detail in numerous subsequent theoretical studies^{7–10}. The electronic properties of non-molecular nitrogen strongly depend on the particular crystalline structure. Many structures have been considered^{3–5}; the lowest-energy modifications reported seems to

be nonmetallic cubic *gauche*⁵ structure or the A7 (α -arsenic) phases which are semimetallic or nonmetallic^{4,6}. The theoretical studies indicated that the insulating (semiconducting) and metallic phases are very close in energy⁹.

The large apparent difference between theory and experiment for the transition pressure has long been a problem in high-pressure physics. Shock-wave experiments carried out on fluid nitrogen revealed a transition at around 30 GPa and 6,000 K (ref. 11), which was ascribed to molecular dissociation induced by the combined effects of pressure and temperature. Early static compression experiments on solid nitrogen, however, provided no direct evidence for the predicted transformation, up to about 150 GPa (refs 12, 13). That could be explained by a very large hysteresis for the transition in the solid, suggested by the large calculated volume and barrier to dissociation change: $\Delta V \approx 35\%$ (refs 3, 4) and the 1 eV per atom change in cohesive energy⁵. However, such a large hysteresis would require a pressure metastability that perhaps exceeds 100 GPa, for which there was no direct evidence, in nitrogen

or any other materials. Clarification of this issue is crucial for assessing the reliability of theoretical calculations^{14,15} for the related transition in hydrogen, and for understanding the kinetics of high-pressure transformations.

We used new ultrahigh-pressure electrical resistance techniques, in combination with Raman spectroscopy, to exert pressures of above 240 GPa. Electrical measurements provide indisputable proof of the existence of a metal or non-metal, whereas optical measurements in a diamond cell cannot examine the optical conductivity to the lowest (that is, zero) frequency needed to distinguish between the two forms. But electrical measurements at megabar pressures currently present an experimental challenge—here, we have extended the accessible pressure region for electrical measurements of condensed gases by almost one megabar. This doubles the pressure range of previous experiments on oxygen to 120 GPa (ref. 16), and very recent measurements on Xe to 155 GPa (ref. 17). We performed 12 runs to maximum pressures of 110, 140, 155, 160, 193 and 240 GPa at various temperatures from 300 K to 80 K, including 7 runs on decompression to examine the kinetics and hysteresis of the transition and the stability of the non-molecular phase at low pressures.

In the first run at 190 GPa at 80 K, for example, the sample suddenly became opaque (Fig. 1). This sudden change is accompanied by an abrupt loss in intensity of the vibron (N–N stretching mode; Fig. 2). Both observations indicate a strong first-order phase transformation. Despite these changes in optical properties, the electrical resistance measurements showed that the sample was insulating, with the resistance remaining above 300 M Ω at all pressures, including above the transition. Continued observation

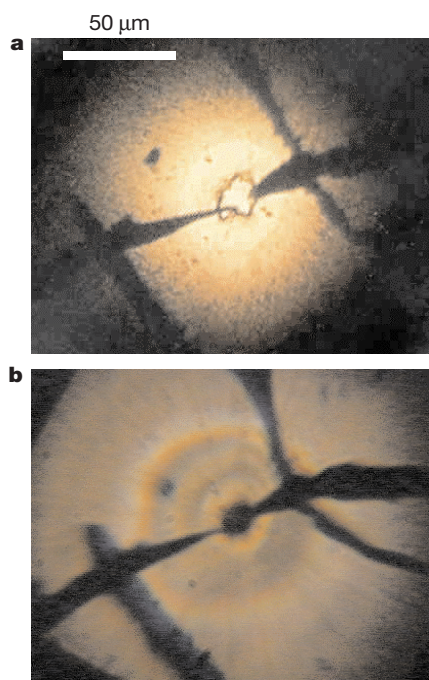


Figure 1 Photomicrographs of the sample of nitrogen at 70 GPa (a) and 193 GPa (b). Nitrogen was confined in a hole of diameter 20 μm in the insulating gasket. The sample showed negligible changes on increase of pressure up to the abrupt transition at 190 GPa. The transition occurred abruptly at 190 GPa, as indicated by an observable contraction and darkening of the sample at the edge of the gasket, and loss of the Raman vibron (Fig. 2). With further loading, the sample suddenly (within minutes) transformed to an opaque phase, indicative of a major phase transition. Pressure was generated by a pair of diamonds with 300- μm culets, a 9° bevel angle, and a 50- μm flat surface at the tip. Ultrahigh-purity nitrogen was loaded as a gas into a 20- μm gasket hole of a megabar-type diamond cell¹⁹ at a pressure of 0.2 GPa. The experimental arrangement was similar to that described in ref. 17. Platinum foil electrodes were insulated by a layer made from a mixture of cubic boron nitride (1- μm powder) and epoxy. This arrangement of electrodes allowed us to use the quasi-four-electrode method for the resistance measurement. Two electrodes were inserted into the sample, and each of these was connected to additional electrodes near it to form a cross (as shown in the image). The direct electrical contacts between the pairs of electrodes allowed us to test whether the electrodes were broken during loading. The visible portions of the electrodes from the cross to the sample were obviously unbroken in each run, as clearly observed under the microscope. In each of the runs, we routinely checked that there was no contact between the electrodes and the gasket. The gasket was shown to be non-conducting, as demonstrated by control experiments on H_2O ice, performed with an identical gasket up to 150 GPa. There was no evidence of chemical reactions between platinum electrodes with nitrogen.

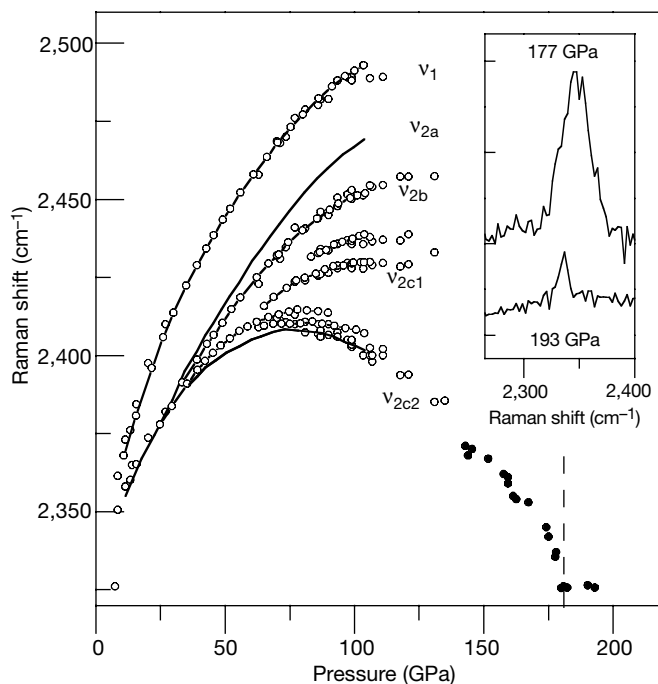


Figure 2 Pressure shifts of the vibron frequencies of nitrogen at 300 K (open circles) and 80 K (solid circles). The lines are the Raman shifts reported by ref. 23. The pressure of the transformation to the new phase observed at 80 K is shown by the vertical dashed line. The inset shows the marked decrease in intensity of the principal vibron. The narrowing of the vibron indicates a significant decrease in pressure gradients over the sample after the transition. The position of the principal vibron levels remains constant (2,325 cm⁻¹) over a range of pressures in the vicinity of the transition owing to phase coexistence (as typically observed in first-order phase transitions). The pressure was measured from ruby chips²⁴ embedded within the nitrogen sample at all temperatures in an optical cryostat. Both Raman and ruby luminescence spectra were measured with visible lines of an Ar⁺ laser and with variable red wavelengths of a Ti-sapphire laser.

of the crossed pairs of electrodes indicated that all four electrodes remained intact under pressure (see Fig. 1). We thus conclude that the black high-pressure phase at the low temperatures is not a metal but an insulator or a semiconductor with a large energy gap. Our subsequent optical studies provided indirect evidence for the semiconducting state in the non-molecular phase¹⁸.

We performed further runs at various temperatures and at higher pressures up to 240 GPa; these runs revealed the semiconducting nature of the new phase. Electrical measurements at 300 K showed an onset of measurable conductivity near 140 GPa. This is accompanied by gradual changes in the optical properties and Raman spectra of the molecular nitrogen, consistent with refs 12, 13, 18 (and R.J.H., unpublished work). A measurable resistance appeared at this pressure. The resistance decreased with pressure (Fig. 3), but remains high to 240 GPa (estimated resistivity of $\sim 100 \Omega \text{ cm}$). The temperature dependence of the resistance indicates that the non-molecular phase is semiconducting (inset of Fig. 3). We attributed an activation energy of 0.4 eV for conductance at high temperature to the intrinsic gap. The alternative interpretation—that the activation energy arises from an impurity level within the intrinsic gap that has a significantly higher energy—contradicts the optical observation of the gap¹⁸ and the optical data obtained in the present experiments. The optical gap ($\sim 0.6 \text{ eV}$ at 170 GPa, decreasing with pressure) is very close to the thermal gap we obtained from electrical measurements. The optical data also show that the gap is intrinsic because absorption is high ($>10^4 \text{ cm}^{-1}$), which is typical for direct interband transitions, and very unlikely for impurity absorption. The change in the activation energy (inset Fig. 3) may, however, be the result of a phase transition, and therefore the 0.18 eV can be ascribed to the energy gap of the low-temperature phase. This

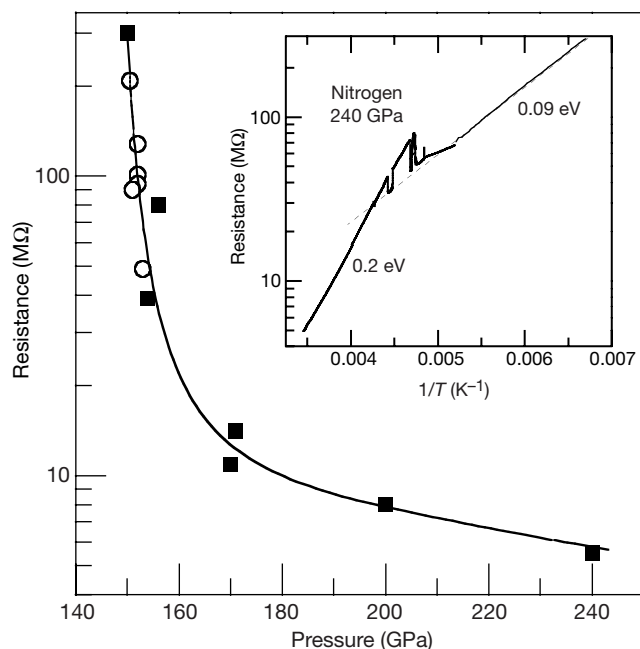


Figure 3 Pressure dependence of resistance of nitrogen at room temperature. The experimental points correspond to two runs (lines are guides to the eye). The inset shows the semiconducting behaviour of nitrogen at 240 GPa. We used standard methods for determining the energy gap: the temperature dependence of the resistance of an intrinsic semiconductor is dominated by the exponential dependence $\exp(-E_g/2k_B T)$ of carrier concentration, where E_g is the energy gap, and k_B is Boltzmann's constant. Therefore, the slope at the temperature dependence ($1/T$) of the resistance ($\ln R$) gives $E_g/2$. The temperature dependence of the resistance of nitrogen plotted in this manner (inset) reveals two linear regimes giving energy gaps of 0.4 eV and 0.18 eV ascribed to two phases at high and low temperatures, respectively (see text).

interpretation is attractive because it could also explain the irregular drops in the temperature dependence of resistance in the transition region (inset Fig. 3): these drops could then arise from switching of current between grains of two mixed phases. Thus, unlike hydrogen, which is predicted to become metallic on dissociation, nitrogen is semiconducting in the post-molecular phase. Moreover, the persistence of the semiconducting phase to such high densities, despite the dissociation of the molecules, is remarkable.

Equally intriguing is the behaviour of the material on decompression. The apparent contradiction between theory and experiment on the stability of the solid molecular phase to very high pressures has been the subject of several theoretical studies: first-principles calculations indicated that the equilibrium transition

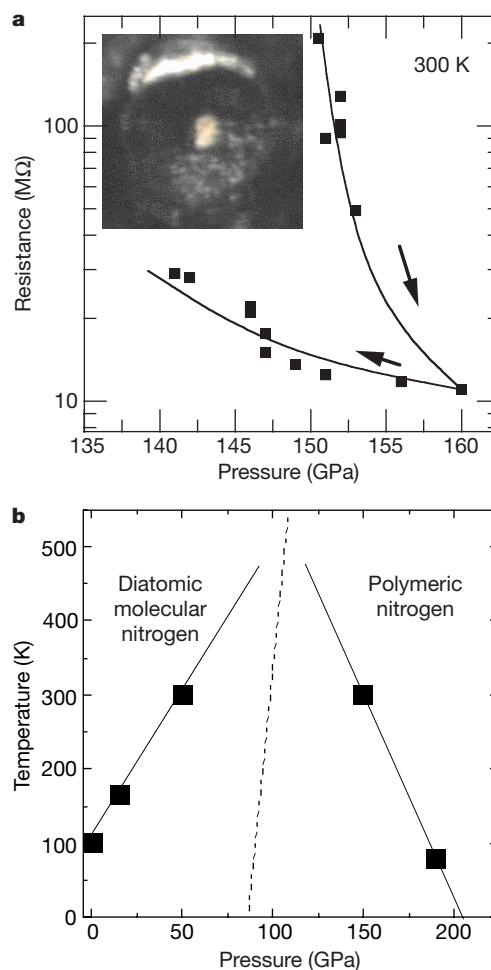


Figure 4 Hysteresis of the molecular–non-molecular nitrogen transition. **a**, Pressure dependence of resistance of nitrogen at room temperature with increasing and decreasing pressure (direction shown by arrows). The line is a guide to the eye. Electrical leads were broken at 140 GPa. With further decrease in pressure the sample gradually became transparent and colourless between 57 and 48 GPa. An intermediate state of the sample at 50 GPa is shown in the inset. The transparent sample revealed Raman spectra characteristic of molecular nitrogen at this pressure. The diameter of the culet (visible circle) is $68 \mu\text{m}$, and the diameter of the sample (red spot) is $12\text{--}15 \mu\text{m}$. **b**, Transformation diagram of nitrogen. The symbols show the experimental points for the forward and back-transformations. The width of the symbol indicates the pressure range over which the transformation was observed to take place. The lines show schematic boundaries for the forward and back-transformations. The equilibrium line (dashed line) was estimated from the midpoint of the transition onsets on compression and decompression.

should occur at 50–94 GPa (refs 3–5), in contrast to the transition at 190 GPa at 80 K and near 150 GPa at 300 K documented experimentally (the present work and ref. 18). The higher pressure of the transition onset at low temperature suggests an even larger discrepancy between experiment and theory (the calculations correspond to $T = 0$ K). If the measured onsets represent an equilibrium transition line, the apparent temperature dependence would indicate a significant negative Clapeyron slope for the transition; this would mean that the high-pressure phase has a lower entropy, possibly consistent with its higher coordination and longer N–N bond distances. On the other hand, the observation of the higher transition pressure at lower temperatures could be a purely kinetic effect, as suggested by theory.

To test these two hypotheses, we studied nitrogen samples decompressed from above 150 GPa at temperatures from 300 K to 10 K. Seven runs were performed in which we pressurized nitrogen to the highest pressures to form the opaque phase, and then released the pressure to check the hysteresis directly. In each case, the non-molecular phase persisted well below 100 GPa. In several runs, the high-pressure nitrogen sample escaped by rupturing of the gasket. This can be explained by the large increase of volume needed for the back-transformation to the molecular phase. In other experiments, however, we succeeded in observing the back-transformation; the following was typical. On releasing pressure from 160 GPa at room temperature, the sample did not initially change visually. However, the resistance of the opaque phase increased (Fig. 4), consistent with the increasing energy gap of the semiconducting phase with decreasing pressure. At 57 GPa, the sample started to become translucent in regions where nucleation of the forward transition had been observed on compression (that is, near pieces of ruby). Gradually within 1 hour, the sample changed from opaque to transparent. At the same time, the pressure decreased to 48 GPa. At the initial stages of the transformation (red colour) the sample showed an enhanced luminescence, indicating enhanced defects that can accompany the change in structure. The luminescence decreased with increasing transparency of the sample. Raman spectra from the colourless sample exhibited vibron peaks characteristic of molecular nitrogen at these pressures.

When the decompression was carried out at lower temperatures, the sample back-transformed at lower pressures. For example, a sample was pressurized to 165 GPa at 300 K and converted to the non-molecular phase. The pressure was then reduced to 90 GPa and the sample cooled to 10 K; at this temperature the pressure was reduced to 15 GPa. When warmed, the sample remained opaque at temperatures up to 175 K, as determined from both visual observations and measurements of the transmission spectrum (giving an absorption edge at about 700 nm). Above this temperature the sample escaped. The results show that the non-molecular phase can be stabilized from 10 K to at least 175 K at 10–20 GPa. Additional evidence for the large hysteresis and high-pressure metastability of the molecular phase was also noted in the compression cycles of these experiments. Specifically, nuclei of the high-pressure phase could be observed at pressures well below that of the overall transformation: opaque areas appeared near defects in the sample (for example, around the ruby chips and near the edges of the gasket), an observation that varied from one sample to the next. Typically, the darkening began at 120 to 130 GPa, but in one run we observed darkening around the ruby at 98 GPa. At a pressure of 150 GPa at 300 K, the whole sample darkened from red to black. At the same time a measurable resistance appeared.

In numerous further attempts, we tried to recover the sample to ambient pressure. In most cases the non-molecular sample escaped by rupturing the gasket at pressures of about 50 GPa because of the inherent weakness of the gasket on unloading¹⁹. However, we did succeed in recovering the non-molecular phase to atmospheric pressure. The sample was completely opaque at low temperatures but became transparent at around 100 K. Properties of the ambient-

pressure non-molecular nitrogen as well as details of the phase transformations including transient and disordered states will be published elsewhere. From these results, we can construct the kinetic and equilibrium phase boundaries for the transition (Fig. 4b). This transition has a surprisingly large hysteresis loop ΔP of 100 GPa at room temperature and $\Delta P > 170$ GPa at 80 K. The equilibrium pressure of this transition is about 100 GPa at room temperature, and is likely to be lower at low temperature, in apparent good agreement with first-principles calculations^{3–5}.

Thus optical and electrical measurements show that nitrogen forms a non-molecular semiconductor that exists at equilibrium over a very large range of pressures, from below 100 GPa to at least 240 GPa. The transformation is associated with an exceedingly large hysteresis loop (ΔP approaching 200 GPa at low temperature). The excellent agreement between theory and experiment for the large hysteresis and slow kinetics of the transition, as well as the determination of the equilibrium phase boundary near 100 GPa supports the accuracy of current first-principles theories for such molecular materials^{3–5}. This work also demonstrates that direct measurements of the electrical resistance of condensed gases can now be performed to ultrahigh pressures well above 200 GPa, making possible studies of other materials such as hydrogen at these conditions. We note that a large hysteresis has been proposed for the still-hypothetical non-molecular (metallic) phase of solid hydrogen, with the possibility of quenching the material to low pressures²⁰. Moreover, the present study suggests the possibility of important hysteresis effects associated with other pressure-induced molecular dissociation transitions, including that predicted for hydrogen, which is found to persist in the molecular state on compression to around 300 GPa (refs 21, 22). Continued study of this unusual form of nitrogen and related materials could result in applications of these dense substances at ambient pressure. □

Received 4 October 2000; accepted 12 March 2001.

- Freiman, Yu. A. in *Physics of Cryocrystals* (eds Manzhelii, V. G. & Freiman, Yu. A.) 538–596 (American Institute of Physics, Woodbury, New York, 1997).
- Wigner, E. & Huntington, H. B. On the possibility of a metallic modification of hydrogen. *J. Chem. Phys.* **3**, 764–770 (1935).
- McMahan, A. K. & LeSar, R. Pressure dissociation of solid nitrogen under 1 Mbar. *Phys. Rev. Lett.* **54**, 1929–1932 (1985).
- Martin, R. M. & Needs, R. Theoretical study of the molecular-to-nonmolecular transformation of nitrogen at high pressures. *Phys. Rev. B* **34**, 5082–5092 (1986).
- Mailhot, C., Yang, L. H. & McMahan, A. K. Polymeric nitrogen. *Phys. Rev. B* **46**, 14419–14435 (1992).
- Christe, K. O., Wilson, W. W., Sheehy, J. A. & Boatz, J. A. N₅⁺: A novel homoleptic polynitrogen ion as a high energy density material. *Angew. Chem. Int. Edn* **38**, 2002–2009 (1999).
- Helmy, A. A. Calculation of the pressure-induced insulator-metal transition of nitrogen. *J. Phys. Condens. Matter* **6**, 985–988 (1994).
- Lewis, S. P. & Cohen, M. L. High-pressure atomic phases of solid nitrogen. *Phys. Rev. B* **46**, 11117–11120 (1992).
- Mitas, L. & Martin, R. M. Quantum Monte Carlo of nitrogen: atom, dimer, atomic, and molecular solids. *Phys. Rev. Lett.* **72**, 2438–2441 (1994).
- Yakub, E. S. Short-range intermolecular interaction and phase transitions with rearrangement of chemical bonds. *J. Low-Temp. Phys.* **111**, 357–364 (1998).
- Nellis, W. J., Holmes, N. C., Mitchell, A. C. & van Thiel, M. Phase transition in fluid nitrogen at high densities and temperatures. *Phys. Rev. Lett.* **55**, 1661–1664 (1984).
- Reichlin, R., Schiferl, D., Martin, S., Vanderborgh, C. & Mills, R. L. Optical studies of nitrogen to 130 GPa. *Phys. Rev. Lett.* **55**, 1464–1467 (1985).
- Bell, P. M., Mao, H. K. & Hemley, R. J. Observations of solid H₂, D₂, N₂ at pressures around 1.5 megabar at 25 K. *Physica B* **139–140**, 16–20 (1986).
- Johnson, K. A. & Ashcroft, N. W. Structure and bandgap closure in dense hydrogen. *Nature* **403**, 632–635 (2000).
- Städele, M. & Martin, R. M. Metallization of molecular hydrogen: predictions from exact-exchange calculations. *Phys. Rev. Lett.* **84**, 6070–6073 (2000).
- Shimizu, K., Suhara, K., Ikumo, M., Eremets, M. I. & Amaya, K. Superconductivity in oxygen. *Nature* **393**, 767–769 (1998).
- Eremets, M. I. et al. Electrical conductivity of Xe at megabar pressures. *Phys. Rev. Lett.* **83**, 2797–2800 (2000).
- Goncharov, A. F., Gregoryanz, E., Mao, H. K., Liu, Z. & Hemley, R. J. Optical evidence for nonmolecular phase of nitrogen above 150 GPa. *Phys. Rev. Lett.* **85**, 1262–1265 (2000).
- Eremets, M. I. *Experimental High-Pressure Techniques* 16, 59 (Oxford Univ. Press, Oxford, 1996).
- Brovman, E. G., Kagan, Y. & Kholas, A. Structure of metallic hydrogen at zero pressure. *Sov. Phys. JETP* **34**, 1300–1315 (1972).
- Hemley, R. J. & Mao, H. K. Phase transition in solid molecular hydrogen at ultrahigh pressures. *Phys. Rev. Lett.* **61**, 857–860 (1988).

22. Nayarana, C., Luo, H., Orloff, J. & Ruoff, A. L. Solid hydrogen at 342 GPa: no evidence for an alkali metal. *Nature* **393**, 46–49 (1998).
23. Olijnyk, H. & Jephcoat, A. P. Vibrational dynamics of isotopically dilute nitrogen to 104 GPa. *Phys. Rev. Lett.* **83**, 332–335 (1999).
24. Mao, H. K., Xu, J. & Bell, P. M. Calibration of the ruby pressure gauge to 800 kbar under quasihydrostatic conditions. *J. Geophys. Res.* **B 91**, 4673–4676 (1986).

Acknowledgements

We thank S. Gramsch, V. V. Struzhkin and A. F. Goncharov for useful discussions and comments on the manuscript. This work was supported by the NSF and NASA.

Correspondence and requests for materials should be addressed to R.J.H. (e-mail: hemley@gl.ciw.edu).

A prototype storage ring for neutral molecules

Floris M.H. Crompvoets, Hendrick L. Bethlem, Rienk T. Jongma & Gerard Meijer

FOM-Institute for Plasma Physics Rijnhuizen, PO Box 1207, NL-3430 BE Nieuwegein, The Netherlands; and the Department of Molecular and Laser Physics, University of Nijmegen, Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands

The ability to cool and manipulate atoms with light has yielded atom interferometry, precision spectroscopy, Bose–Einstein condensates and atom lasers. The extension of controlled manipulation to molecules is expected to be similarly rewarding, but molecules are not as amenable to manipulation by light owing to a far more complex energy-level spectrum. However, time-varying electric and magnetic fields have been successfully used to control the position and velocity of ions, suggesting that these schemes can also be used to manipulate neutral particles having an electric or magnetic dipole moment^{1–4}. Although the forces exerted on neutral species are many orders of magnitude smaller than those exerted on ions, beams of neutral dipolar molecules have been successfully slowed down in a series of pulsed electric fields^{5,6} and subsequently loaded into an electrostatic trap⁷. Here we extend the scheme to include a prototype electrostatic storage ring made of a hexapole torus with a circumference of 80 cm. After

injection, decelerated bunches of deuterated ammonia molecules, each containing about 10^6 molecules in a single quantum state and with a translational temperature of 10 mK, travel up to six times around the ring. Stochastic cooling⁸ might provide a means to increase the phase-space density of the stored molecules in the storage ring, and we expect this to open up new opportunities for molecular spectroscopy and studies of cold molecular collisions^{9,10}.

The experimental set-up (Fig. 1) consists of a compact molecular beam machine that tangentially injects molecules into a hexapole torus storage ring. The centripetal force required to keep the molecules in a circular orbit arises from the interaction of the molecules with the inhomogeneous electric field inside the ring. For a given maximum electric-field strength, a sufficiently large centripetal force can be generated only in a storage ring with a minimum radius that is proportional to the square of the velocity of the injected molecules. The availability of beams of slow neutral molecules, with a tunable absolute velocity^{5,11}, is therefore an important asset in designing and testing a molecular storage ring. To simplify detection, the electric field geometry in the 25-cm-diameter hexapole torus storage ring deviates from the commonly used cylindrically symmetric hexapole fields. Figure 2 illustrates the voltages applied to the different rods of the ring and the equipotential lines in the shallow well experienced by state-selected deuterated ammonia molecules travelling with a tangential velocity of 89 m s^{-1} .

A pulsed beam of deuterated ammonia, ND_3 , is produced by expanding a ND_3/Xe mixture through a modified solenoid valve into vacuum. Seeding ND_3 in the heaviest rare gas and cooling of the pulsed valve results in a mean beam velocity of around 275 m s^{-1} ($E_{\text{kin}} = 64 \text{ cm}^{-1}$). About 15% of the ND_3 molecules reside in the low-field-seeking $|J K| M\rangle = |1 1 1\rangle$ upper inversion level of the vibrational ground state. We select only molecules that are in this quantum state, molecules that gain Stark energy with increasing electric field and are repelled from high-electric-field regions. The electric dipole moment of ND_3 is 1.5 D (Debye units)¹², and the selected molecules experience a positive Stark shift of 1.2 cm^{-1} in an electric field of 100 kV cm^{-1} . By selecting ND_3 , rather than NH_3 , we ensure that we can work with a more linear Stark effect in the applied electric fields, resulting from the fact that the magnitude of the inversion splitting is significantly reduced upon deuteration¹².

The ND_3 molecules pass through a 1.0-mm-diameter skimmer into a second, differentially pumped vacuum chamber and then fly into a short pulsed hexapole that acts as a positive lens for the selected ND_3 molecules. After exiting the hexapole, the molecules

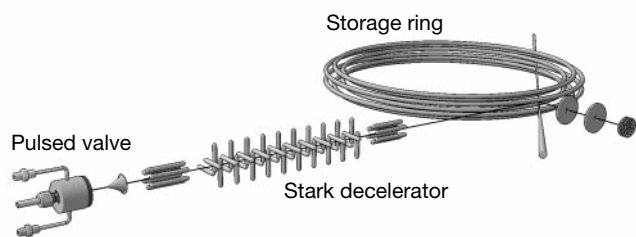


Figure 1 Experimental set-up. From left to right: a gas pulse of less than 1% ND_3 in Xe exits a cooled ($T = 200 \text{ K}$) pulsed valve with a mean velocity of 275 m s^{-1} . After passing through the skimmer the beam is collimated with a pulsed hexapole and subsets of these molecules are slowed down to velocities in the 76 m s^{-1} to 110 m s^{-1} range upon passage through the Stark decelerator. The bunches of slow molecules are tangentially injected into the electrostatic storage ring (with a 12.5-cm radius and 4-mm-diameter rods). Molecules in the detection region of the storage ring are ionized using a pulsed laser, after which they are extracted, and detected with an ion detector. The flight path of the molecules from the pulsed valve to the detection region in the storage ring is 65 cm.

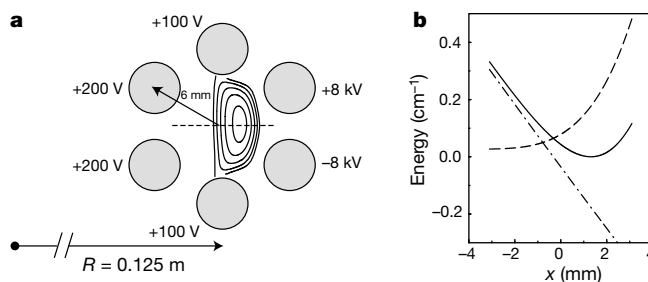


Figure 2 Cross-section of the hexapole storage ring. **a**, Contour lines of equal potential energy for ND_3 molecules in the $|J K| M\rangle = |1 1 1\rangle$ state with positive Stark shift and with a tangential velocity of 89 m s^{-1} are shown at 0.02 cm^{-1} intervals. **b**, The dashed curve shows the Stark energy as a function of the displacement from the centre of the hexapole (x) in the plane of the hexapole ring (along the dashed axis shown in **a**). The solid curve is obtained by adding the potential energy as a result of the centrifugal force for an ammonia molecule with a tangential velocity of 89 m s^{-1} (dot-dashed line) to the Stark energy (dashed line). The minimum of the resulting potential well is located 1.3 mm offset from the geometrical centre of the hexapole.

enter a 35-cm-long Stark decelerator containing an array of equidistant electric-field stages, with electric fields of typically up to 100 kV cm^{-1} ; this decelerates the molecules while maintaining their initial phase-space density^{5,6}, much like what is done in charged-particle accelerators¹³.

A second pulsed hexapole at the exit of the decelerator focuses the slowed-down molecules into the storage ring. The ammonia molecules enter the ring tangentially, passing through the 2-mm-wide 'slit' between two hexapole rods. The density of ND_3 molecules inside the storage ring is probed using the focused radiation of a pulsed laser at 317 nm, which ionizes (by a multi-photon process) the ND_3 molecules that are in the selected quantum state¹⁴. Just before firing the laser, the voltages on the hexapole ring are switched such that laser-produced positive ions are accelerated radially outward, so that the storage ring itself serves as the extraction region of a linear time-of-flight mass spectrometer. The ND_3 -ion signal is recorded and used as a measure for the amount of neutral ammonia molecules originally present inside the ring.

Figure 3 illustrates the relative density of neutral ammonia inside the storage ring, as a function of the time after the molecules left the pulsed valve. Without any voltage applied to the Stark decelerator (that is, without deceleration), the ammonia density in the storage ring peaks around 2.4 ms after the molecules left the valve. This is the time it takes molecules travelling with an average velocity of 275 m s^{-1} to cover the 65-cm distance between the pulsed valve and the storage ring. The velocity spread is approximately 20% (full-width at half-maximum, FWHM), implying a translational temperature of around 1 K in the moving frame. With the Stark decelerator switched on and configured to decelerate molecules from 275 m s^{-1} to a final velocity of 100 m s^{-1} , a series of signal peaks occur at later times. These different bunches, whose velocity is indicated in Fig. 3, originate from the spatially extended molecular gas pulse at the entrance of the decelerator. The molecules that exit the decelerator with a velocity of 110 m s^{-1} were already in the second electric-field section of the decelerator when the decelerator was switched on, and so missed one deceleration cycle. The molecules that exit the decelerator with velocities of 89 m s^{-1} and 76 m s^{-1} are molecules that entered the decelerator later, with an initial velocity that was lower than the average beam velocity. The individual bunches are still clearly separated in the probe region,

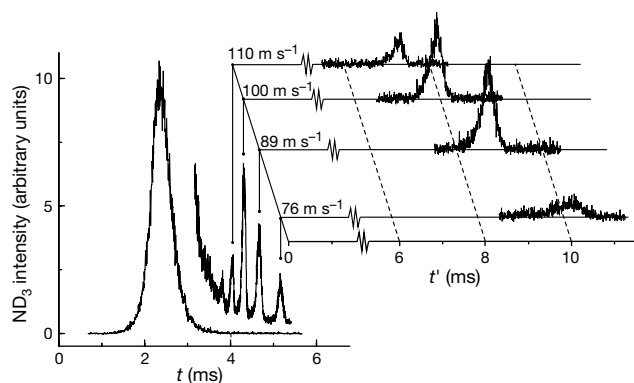


Figure 3 Time-of-flight profiles. Density of ammonia molecules inside the storage ring as a function of time after opening of the pulsed valve (t), both without (single large peak around 2.4 ms) and with the Stark decelerator on (series of peaks at later arrival times). It is worth noting that the intensity of the decelerated bunches is on the same absolute scale as the original beam; the focusing effects in the decelerator typically enhance the signal on the beam axis by a factor of 5 compared to the situation when the decelerator is not switched on at all. In the inset the density of ammonia molecules inside the storage ring is measured as a function of the time after switching on the storage ring (t'). Bunches of decelerated molecules with four different centre velocities are injected into the ring, with settings as shown in Fig. 2, and their time-of-flight distributions after the first round trip are recorded.

which is located 19 cm away from the exit of the decelerator. The velocity spread in the decelerated bunches of molecules is only 4–5 m s^{-1} (FWHM), corresponding to a translational temperature of around 10 mK. The detection signal we obtain corresponds to about a hundred ND_3 -ions being generated per laser pulse. As ionization of the ND_3 molecules takes place only in the focus of the laser beam, the actual detection volume in our system is about 10^{-4} cm^3 . Each bunch contains a minimum of about 10^6 state-selected ND_3 molecules.

When a selected bunch of decelerated ammonia molecules has entered the storage ring, a high voltage is abruptly applied to the outer two hexapole rods. The separation between the different bunches at the entrance region of the storage ring is sufficiently large that only one selected bunch of decelerated molecules is captured. To be able to detect the amount of ammonia molecules inside the ring after having stored them for a certain time, the voltages on the outer two rods are switched off; the (small) voltage offset required for ion extraction is permanently applied to the other rods. The inset to Fig. 3 shows the density of neutral ammonia in the detection region of the storage ring as a function of time after switching on the ring. After separately injecting each of the four selected bunches of molecules, the corresponding time-of-flight distributions after one round trip are measured. The measurements show that ammonia molecules in the velocity interval from 76 m s^{-1} to 110 m s^{-1} can be captured in the ring with the settings as applied. Molecules that move significantly faster do not experience a force large enough to keep them in a circular orbit, whereas molecules that move significantly slower will be deflected too much by the repulsive outer wall of the ring, and either hit the inner hexapole rods or escape from the storage ring on the inside otherwise. With the settings that have been used, the injection efficiency is optimum for velocities around 89 m s^{-1} .

The equilibrium orbit for molecules with a forward velocity of 89 m s^{-1} coupled in at the minimum of the potential well is a perfect circle in the horizontal plane, with a radius 1.3 mm larger than the 12.5 cm radius of the hexapole torus. Upon making successive round trips the bunch of molecules gradually spreads out in the tangential direction as a result of the residual velocity spread. In addition, the molecules will perform betatron oscillations around the equilibrium orbit¹³. In our particular geometry the oscillations in the radial and vertical direction are strongly coupled. Per round trip, approximately eight oscillations are made in the radial direction

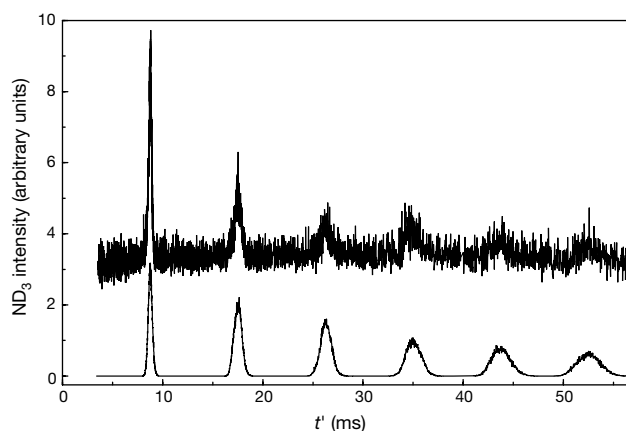


Figure 4 Multiple round trips. The measurements shown in the upper trace show that bunches containing approximately a million ammonia molecules make up to six round trips in the storage ring. The package of stored molecules is seen to gradually broaden and to decrease in peak intensity, mainly due to the tangential velocity spread (dispersion). In the lower trace the results of Monte Carlo simulations, calculating trajectories of molecules through the whole experimental set-up, are shown. From the simulations a translational temperature of 10 mK is deduced for the stored bunch of molecules.

and three in the vertical direction. Conservation of angular momentum dictates that the tangential velocity of the orbiting molecules is modulated at the radial betatron oscillation frequency as well. The experimental data in the upper trace of Fig. 4 demonstrates that a bunch of ammonia molecules with a velocity of 89 m s^{-1} can still be identified after it has made six round trips in the storage ring. The observed gradual broadening and decrease in peak intensity after each additional round trip is largely explained by the above-mentioned spreading out of the package along the ring, as substantiated by the results from Monte Carlo simulations, shown in the lower trace of Fig. 4. Loss of molecules from the storage ring by collisions with background gas can explain the observed faster decay of signal at early times after the gas pulse, when the background pressure is still rather high. Unwanted transitions to states that are degenerate in zero field—another possible loss channel for the ring—are avoided in the chosen electric-field geometry.

The storage ring experimentally demonstrated here can generally be used to confine dipolar molecules. The advantage of a storage ring over a trap is that bunches of cold molecules in a ring can be made to interact repeatedly, at well defined times and at distinct locations, with electromagnetic fields and/or other particles. A sectional version of the storage ring allows implementation of field-free interaction regions and various out-coupling regions. This yields unique opportunities for high-resolution molecular spectroscopy as well as for a large variety of collision studies. By incorporating a re-bunching element in the ring, the gradual spreading out of the package that is observed now can be avoided and the stored molecules can be confined to the same region in phase-space for a time-period only limited by collisions with background gas. Different molecules can be stored simultaneously; slowly overtaking or counter-propagating bunches can be stored and used for collision studies. The excellent optical access to the storage ring makes it feasible to use laser detection schemes to accurately determine where the molecules are in the ring at each given moment in time. If this is done in a non-intrusive manner, for instance via two-level laser-induced fluorescence or by detecting the change in refractive index when they pass through an optical cavity, this information can be used to actively correct the trajectories of individual molecules towards the ideal, circular orbit, thus increasing the phase-space density of the stored molecules (stochastic cooling⁸). It is important to note that in the ring, molecules in high-field-seeking states can also be stored¹⁵; motional stabilization as demonstrated here can also be used to stabilize orbits of particles that, according to the Earnshaw theorem, cannot be trapped in a static geometry. Whenever any cooling scheme on the stored molecules is proved to work (stochastic cooling, evaporative cooling, laser cooling on ro-vibrational transitions, cavity-mediated cooling¹⁶), reloading of the ring will be possible. Such reloading might be a viable way of achieving phase-space densities of stored molecules that are sufficiently high to permit experimental study of molecular quantum-collective effects. □

Received 15 January; accepted 23 March 2001.

- Küger, K.-J., Paul, W. & Trinks, U. A magnetic storage ring for neutrons. *Phys. Lett. B* **72**, 422–424 (1978).
- Niel, L. & Rauch, H. Acceleration, deceleration and monochromatization of neutrons in time dependent magnetic fields. *Z. Phys. B* **74**, 133–139 (1989).
- Huffman, P. R. *et al.* Magnetic trapping of neutrons. *Nature* **403**, 62–64 (2000).
- Weinstein, J. D., deCarvalho, R., Guillet, T., Friedrich, B. & Doyle, J. M. Magnetic trapping of calcium monohydride molecules at millikelvin temperatures. *Nature* **395**, 148–150 (1998).
- Bethlem, H. L., Berden, G. & Meijer, G. Decelerating neutral dipolar molecules. *Phys. Rev. Lett.* **83**, 1558–1561 (1999).
- Bethlem, H. L., Berden, G., van Roij, A. J. A., Crompvoets, F. M. H. & Meijer, G. Trapping neutral molecules in traveling potential well. *Phys. Rev. Lett.* **84**, 5744–5747 (2000).
- Bethlem, H. L. *et al.* Electrostatic trapping of ammonia molecules. *Nature* **406**, 491–494 (2000).
- van der Meer, S. Stochastic damping of betatron oscillations in the ISR. CERN internal report CERN/ISR-PO/72-31 (CERN, Geneva, 1972).
- Katz, D. P. A storage ring for polar molecules. *J. Chem. Phys.* **107**, 8491–8501 (1997).
- Weiner, J., Bagnato V. S., Zilio, S. & Julienne, P. S. Experiments and theory in cold and ultracold collisions. *Rev. Mod. Phys.* **71**, 1–85 (1999).

- Gupta, M. & Herschbach, D. A mechanical means to produce intense beams of slow molecules. *J. Phys. Chem. A* **103**, 10670–10673 (1999).
- Gandhi, S. R. & Bernstein, R. B. Focusing and state selection of NH_3 and OCS by the electrostatic hexapole via first- and second-order Stark effects. *J. Chem. Phys.* **87**, 6457–6467 (1987).
- Lee, S. Y. Accelerator physics. (World Scientific, Singapore, 1999).
- Ashfold, M. N. R., Dixon, R. N., Little, N., Stickland, R. J. & Western, C. M. The $\tilde{B}'E''$ state of ammonia: sub-Doppler spectroscopy at vacuum ultraviolet energies. *J. Chem. Phys.* **89**, 1754–1761 (1988).
- Loesch, H. J. & Scheel, B. Molecules on Kepler orbits: an experimental study. *Phys. Rev. Lett.* **85**, 2709–2712 (2000).
- Vuletić, V. & Chu, S. Laser cooling of atoms, ions, or molecules by coherent scattering. *Phys. Rev. Lett.* **84**, 3787–3790 (2000).

Acknowledgements

This work is part of the research programme of the Stichting voor Fundamenteel Onderzoek der Materie (FOM), which is financially supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). The research of R.T.J. has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. We acknowledge the expert technical assistance of A.J.A. van Roij, and the help of G. Berden and J. Ye in the experiments. G. M. dedicates this work to P. Andresen.

Correspondence and requests for information should be addressed to G.M. (e-mail: gerardm@rijnh.nl).

Evidence against dust-mediated control of glacial–interglacial changes in atmospheric CO_2

B. A. Maher*† & P. F. Dennis*

* Centre for Environmental Magnetism and Palaeomagnetism (CEMP) and Stable Isotope Laboratory, School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK

The low concentration of atmospheric CO_2 inferred to have been present during glacial periods is thought to have been partly caused by an increased supply of iron-bearing dust to the ocean surface¹. This is supported by a recent model² that attributes half of the CO_2 reduction during past glacial stages to iron-stimulated uptake of CO_2 by phytoplankton in the Southern Ocean. But atmospheric dust fluxes to the Southern Ocean, even in glacial periods, are thought to be relatively low and therefore it has been proposed that Southern Ocean productivity might be influenced by iron deposited elsewhere—for example, in the Northern Hemisphere^{3,4}—which is then transported south via ocean circulation (similar to the distal supply of iron to the equatorial Pacific Ocean^{5–7}). Here we examine the timing of dust fluxes to the North Atlantic Ocean, in relation to climate records from the Vostok ice core in Antarctica around the time of the penultimate deglaciation (about 130 kyr ago). Two main dust peaks occurred 155 kyr and 130 kyr ago, but neither was associated with the CO_2 rise recorded in the Vostok ice core. This mismatch, together with the low dust flux supplied to the Southern Ocean, suggests that dust-mediated iron fertilization of the Southern Ocean did not significantly influence atmospheric CO_2 at the termination of the penultimate glaciation.

The Southern Ocean is a 'high nutrient, low chlorophyll' (HNLC) region; phosphate and nitrate concentrations are high at the ocean surface yet its productivity is low. Recent experiments show that artificial additions of iron to ocean patches significantly enhance primary productivity⁸, with consequent drawdown of surface water

† Present address: CEMP, Lancaster Environment Centre, Department of Geography, University of Lancaster, LA1 4YP, UK.

f_{CO_2} : the water becomes a sink for atmospheric CO_2 . Hence, the strength of the 'biological pump' may influence atmospheric CO_2 concentrations. The Vostok (Antarctic) ice-core record shows that towards the end of past glacial stages, the Southern Ocean region warmed and atmospheric CO_2 levels rose several thousand years before the Northern Hemisphere ice sheets melted. Thus, this region may greatly influence and/or amplify climate change. A major control on iron supply to HNLC regions is thought to be the flux of terrestrial dust^{1,2}. A model² of the ocean-atmosphere carbon cycle, which predicts up to 40 p.p.m. CO_2 drawdown in response to increased Southern Ocean dust flux during glacial stages, used a modelled⁹ estimate of modern dust flux of $14 \text{ mg cm}^{-2} \text{ kyr}^{-1}$ (a value considered over-estimated⁹) and a Last Glacial Maximum (LGM) value of $345 \text{ mg cm}^{-2} \text{ kyr}^{-1}$. Data on the present-day dust flux to the Southern Ocean, although sparse, are mutually consistent^{10–12} and indicate much lower fluxes of $0.1–1 \text{ mg cm}^{-2} \text{ kyr}^{-1}$. At glacial times, increased aridity and/or wind strength could have contributed to higher Southern Ocean dust fluxes and thus to lower atmospheric CO_2 concentrations. The Vostok record indeed shows peak dust flux during glaciations, when atmospheric CO_2 was depressed by about 80 p.p.m. (ref. 13). However, geological data indicate that these increased glacial dust fluxes were small, and negligible compared with detrital (non-aeolian) sediment accumulation rates in the

Southern Ocean. The Southern Ocean exhibits sedimentation rates far greater than the atmospheric flux, owing to bottom-water transport of detritus. For Scotia Sea sediments, for example, the aeolian contribution from the South American loess regions (identified isotopically as the principal source of Vostok's dust¹⁴) is below the level of detection, even during glacial periods¹⁵. Similarly, in cores from the southeast Indian Ocean to the Antarctic continental slope, aeolian particles account for less than 10% of the clay minerals present¹⁶. An aeolian flux 600 times greater than that actually recorded in the Vostok ice core ($\sim 2 \text{ mg cm}^{-2} \text{ kyr}^{-1}$) would be required to explain the lithogenic flux at these sites during glacial times¹⁶. Thus, the interpretation of lithogenic accumulation rates as a proxy for Southern Ocean dust flux¹⁷ seems to be invalid. Similarly, dust concentration changes in the Vostok ice core cannot be used to constrain model simulations of past dust fluxes² because snow accumulation rates were lower during glacial stages; hence dust flux changes were much smaller than concentration changes⁹.

Given the low dust fluxes to this region, iron supply from upwelling water seems to be a more significant source, by at least an order of magnitude^{3,4}. Similar upwelled iron supply has been identified for the HNLC region of the equatorial Pacific^{5–7}. Thus, for the Southern Ocean, upwelled iron and resultant productivity may

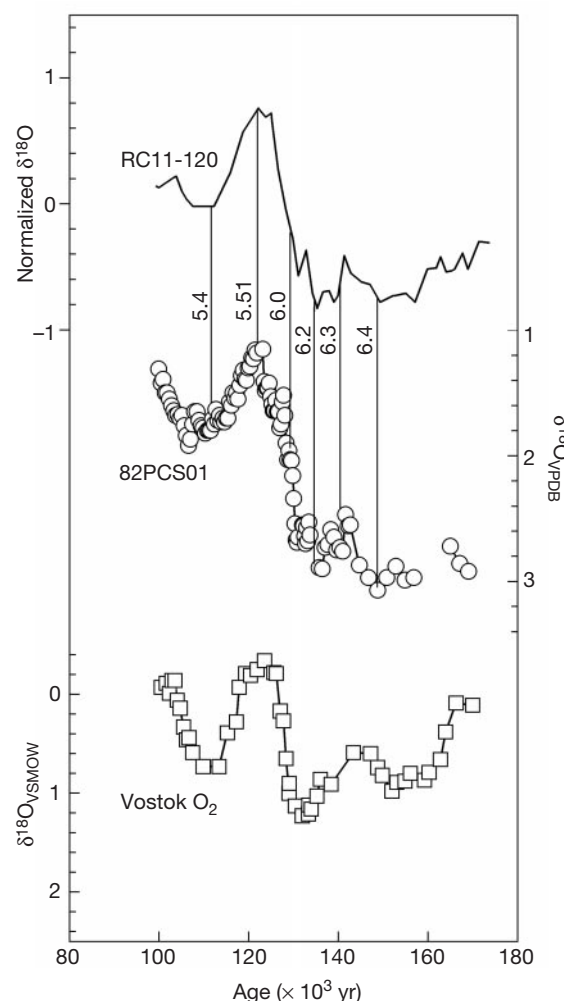


Figure 1 The high-resolution foraminiferal (*Globorotalia bulloides*) oxygen isotope record for piston core 82PCS01 ($42^{\circ}05' \text{ N } 23^{\circ}31' \text{ W}$, water depth 3,540 m). The chronology is based on the indicated tie points between the measured data and the orbitally tuned deep sea record of core RC11-120 (ref. 21). The oxygen isotope substages are marked on the tie lines. The Vostok $\delta^{18}\text{O}$ (in air)¹³, measured with respect to Vienna Standard Mean

Ocean Water (VSMOW), is plotted using the GT4 timescale. For the 100–170 kyr period represented by these records, the chronology is in excellent agreement with an independent age model based on orbital tuning of both the marine and Vostok records to tilt and precession cycles²².

reflect distal (that is, Northern Hemisphere), rather than Southern Hemisphere, dust inputs. It has been suggested that iron from the peri-Saharan region (the largest dust source in the world) may be transported in southward flow of the North Atlantic intermediate/deep water to the circumpolar region, on a timescale of approximately 100 yr (refs 3, 4). This hypothesis suggests that the pattern of Northern, rather than Southern, Hemisphere dust flux may match more closely the pattern and timing of CO₂ variations recorded by the Vostok ice core.

Here we test this hypothesis by high-resolution oxygen isotope, magnetic and elemental analysis of a deep-sea sediment record, core 82PCS01, from the abyssal plain northeast of the Azores. Sedimentation at this site has been continuous and undisturbed from oxygen isotope stage (OIS) 8 to the present day¹⁸. The core site lies above the lysocline, precluding carbonate dissolution, and underlies the present northern margin of the Bermuda–Azores high-pressure system, around which African dust can be transported northwards in the summer¹⁹. Upon sinking, possible complexation or dissolution²⁰, iron may become entrained within southward-flowing Atlantic intermediate/deep water. We obtained an age model for the core via oxygen isotope analysis of the planktonic foraminifer, *Globorotalia bulloides*, correlating the data with the chronostratigraphy of ref. 21. We measured the magnetic remanence acquired at high applied magnetic fields (the HIRM) of the sediments, as a proxy for iron. Additional analyses on subsets of samples included carbonate content, elemental iron and dry bulk density. From these data, we calculated sedimentation rates and fluxes of iron and dust.

Figure 1 shows the oxygen isotope data for core 82PCS01, with the chronology based on tie points to the high-resolution chronostratigraphy²¹. Also shown in Fig. 1 is the Vostok $\delta^{18}\text{O}$ (in air) record¹³. Both these chronologies agree strongly (to within 1 kyr) with the one newly derived by tuning the benthic marine $\delta^{18}\text{O}$ record (and the Vostok air $\delta^{18}\text{O}$ record) to tilt and precession cycles²². Figure 2 shows the $\delta^{18}\text{O}$ and HIRM data for core 82PCS01, for the time interval 100–170 kyr ago. For comparison, terrigenous percentage data are also shown for another, more southerly core in the eastern Atlantic, core 82PCS01 appears to record a representative (but much higher resolution) signal of dust input over at least a regional scale. For glacial stage 6, core 82PCS01 displays two peaks in HIRM, at approximately 130 kyr and ~155 kyr ago. For the succeeding interglacial stage, from about 125 kyr to 100 kyr ago, HIRM values are low and less variable. These high-resolution data amplify those reported²³ for gravity cores in this area. Goethite, a significant component of soils within the semi-arid zone, has been identified as the major contributor to the HIRM of these sediments²³. Given the saturation remanence of goethite²⁴ ($\sim 0.05 \text{ A m}^2 \text{ kg}^{-1}$), our carbonate-free HIRM values ($7\text{--}44 \times 10^{-4} \text{ A m}^2 \text{ kg}^{-1}$) indicate goethite concentrations in the dust component of the sediment of 1.4–8.8%. HIRM values (for every sample) and independently determined iron (in wt%, determined for a subset of samples) are directly and strongly correlated ($R^2 = 0.83$). Thus, rapid, non-destructive HIRM measurements appear to be a robust proxy for aeolian, elemental iron concentrations in these North Atlantic sediments. It should be noted that our North Atlantic record shows particulate iron deposition through time, rather than dissolved iron. The relationship between iron-bearing dust flux and dissolved iron concentrations in deep water is unknown, but dust-sourced increases in deep-water iron would imply increased concentrations of organic ligands capable of complexing the iron to maintain solubility²⁰.

Figure 3a shows the calculated dust fluxes for core 82PCS01, for the period spanning glacial stage 6, Termination II and interglacial stage 5. Dust flux was 3–5 times higher during two intervals: about 155 kyr ago ($\sim 5 \text{ g cm}^{-2} \text{ kyr}^{-1}$) and about 130 kyr ago ($\sim 3 \text{ g cm}^{-2} \text{ kyr}^{-1}$). During interglacial stage 5 (from ~130 kyr ago on), dust fluxes decreased rapidly before steadying at about $1 \text{ g cm}^{-2} \text{ kyr}^{-1}$. For

comparison, estimates of modern fluxes⁹ for the northeast equatorial Atlantic mostly range from $0.2\text{--}3 \text{ g cm}^{-2} \text{ kyr}^{-1}$, with (depending on latitude) a one- to four-fold increase in dust flux at the Last Glacial Maximum.

The oxygen isotope records from 82PCS01 and Vostok are in excellent agreement (Fig. 1 and Fig. 3d and e). At Termination II, the rise in $\delta^{18}\text{O}$ (in air) at Vostok lags the North Atlantic rise by about 2 kyr; this delay is attributable to the Dole effect²². The major, five-fold increase in our North Atlantic dust flux is thus virtually synchronous with peak A in the Vostok dust concentration record (Fig. 3c). However, at its peak, the Vostok dust flux is about 2,500 times smaller than our North Atlantic flux ($\sim 2 \text{ mg cm}^{-2} \text{ kyr}^{-1}$ compared with $5 \text{ g cm}^{-2} \text{ kyr}^{-1}$). In global climate terms, this large rise and fall in North Atlantic dust (and iron) flux occurred entirely within OIS 6, with little apparent variation in either atmospheric CO₂ or oxygen isotope values (Fig. 3d–g). North Atlantic dust flux then declined, substantially earlier (by ~7,000 yr) than the onset of the CO₂ rise at Vostok. The second dust flux event in core 82PCS01, centred on about 130 kyr ago, may have provided a significant, later supply of iron to drive Southern Ocean productivity, when Southern Hemisphere dust flux had diminished to near-zero. However, as shown in Fig. 3, this subsidiary peak in dust flux significantly post-dates (by ~7,000 years) the onset of the CO₂ rise. Rather than being associated with draw-down of CO₂ and maintenance of glacial conditions, the peak of this later event coincides with a marked, 1‰ shift in deep-sea oxygen isotope values (Fig. 3d), indicating either

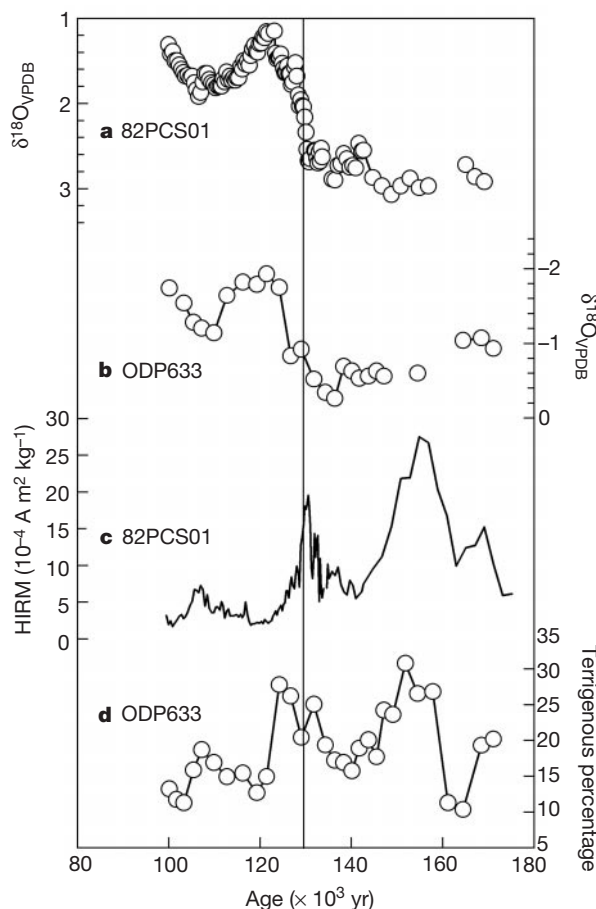


Figure 2 Oxygen isotope and dust (HIRM and terrigenous percentage) records from the eastern North Atlantic. **a**, Oxygen isotope data for piston core 82PCS01, 8.8 m of carbonate ooze interbedded with marly ooze; **b**, oxygen isotope data for Ocean Drilling Program hole 633 (ref. 27), from the eastern tropical Atlantic (1.2° N 11.9° W); **c**, HIRM data for core 82PCS01; **d**, published terrigenous percentage values for ODP hole 663 (ref. 27). The vertical line represents Termination II at about 130 kyr ago²¹.

warming North Atlantic sea surface temperatures²⁵ and/or melting of the Northern Hemisphere ice sheets.

Thus, although fluxes of Northern Hemisphere dust may constitute an alternative, higher-magnitude supply of iron to the Southern Ocean, they seem to bear no causal relationship with either changes in Southern Ocean temperature or atmospheric CO₂, as recorded at Vostok around Termination II. These data—the evidence of very low dust fluxes to the Southern Ocean, and the mismatch of timing between the Northern Hemisphere flux peaks and Southern Ocean climate change—thus do not

appear to support the suggested role of dust-mediated iron fertilization in the Southern Ocean at and around the Termination II boundary. □

Methods

From each 1-cm sample (sieved at 63 µm), 10–20 adult specimens (0.2–0.5 mg) of the planktonic foraminifer, *Globorotalia bulloides*, were picked for oxygen isotope analysis. These were reacted off-line with 102% orthophosphoric acid at 25 °C for 12 h; the evolved CO₂ was dried by passing through a cold trap at –90 °C and cryodistilled into gas sample tubes. Isotopic analyses were performed on a VG SIRA series II dual-inlet isotope ratio mass spectrometer. Oxygen isotopic compositions, with respect to the Vienna Pee-Dee Belemnite (VPDB) standard, are accurate and precise to better than 0.1‰. Analysis of 0.2-mg aliquots of the internal laboratory standard run at random with the foraminiferal samples, at an average frequency of 1 in 4 samples, have a precision ($1\sigma_{n-1}$) of $\pm 0.06\text{‰}$ ($n = 30$). For magnetic analysis, the fraction of each sample less than 63 µm was dried and packed into 10-cm³ plastic cylinders. Magnetic fields were generated using a pulse magnetizer (up to 100 mT) and electromagnet (up to 1 T); magnetic remanences were measured using a fluxgate magnetometer (sensitivity $\approx 10^{-7}$ A m²). Carbonate content was obtained by gasometry, and elemental iron content by X-ray fluorescence of sample beads, fused with sodium lithoborate.

To calculate the iron flux F_{Fe} the following equation was used: $F_{Fe} = r_s \rho_{db} p$, where flux is in units of g cm⁻² kyr⁻¹ and r_s = sedimentation rate (cm kyr⁻¹), ρ_{db} = dry bulk density (g cm⁻³) and p is the weight fraction of iron.

From the age model, a sedimentation rate was calculated for each sediment depth. Heavy sampling meant it was not possible to obtain a ρ_{db} value for every sample depth. Density was estimated²⁶ from the relationship between CaCO₃ content and the measured ρ_{db} of a subset of samples (14) spanning the CaCO₃ range. For values of p , the weight fraction of iron was estimated for every sample from the HIRM/elemental iron correlation. Dust flux was calculated from the p value, assuming an average iron abundance in mineral aerosols of 3.5% (ref. 10).

Received 23 August 2000; accepted 5 March 2001.

- Martin, J. H. Glacial-interglacial CO₂ change: the iron hypothesis. *Paleoceanography* **5**, 1–13 (1990).
- Watson, A. J., Bakker, D. C. E., Ridgwell, A. J., Boyd, P. W. & Law, C. S. Effect of iron supply on Southern Ocean CO₂ uptake and implications for glacial atmospheric CO₂. *Nature* **407**, 730–733 (2000).
- Watson, A. J. & Lefevre, N. The sensitivity of atmospheric CO₂ concentrations to input of iron to the oceans. *Tellus* **B 51**, 453–460 (1999).
- Lefevre, N. & Watson, A. J. Modelling the geochemical cycle of iron in the oceans and its impact on atmospheric carbon dioxide concentrations. *Glob. Biogeochem. Cycles* **13**, 727–736 (1999).
- Loubere, P. Marine control of biological production in the eastern equatorial Pacific Ocean. *Nature* **406**, 497–500 (2000).
- Coale, K. H., Fitzwater, S. E., Gordon, R. M., Johnson, K. S. & Barber, R. T. Control of community growth and export production by upwelled iron in the equatorial Pacific Ocean. *Nature* **379**, 621–624 (1996).
- Wells, M. L., Vallis, G. K. & Silver, E. A. Tectonic processes in Papua New Guinea and past productivity in the eastern equatorial Atlantic. *Nature* **398**, 601–604 (1999).
- Boyd, P. W. *et al.* A mesoscale phytoplankton bloom in the polar Southern Ocean stimulated by iron fertilization. *Nature* **407**, 695–702 (2000).
- Mahowald, N. *et al.* Dust sources and deposition during the last glacial maximum and current climate: A comparison of model results with paleodata from ice cores and marine sediments. *J. Geophys. Res.* **104**, 15895–15916 (1999).
- Duce, R. & Tindale, N. W. Atmospheric transport of iron and its deposition in the ocean. *Limnol. Oceanogr.* **42**, 405–418 (1991).
- Husar, R. B., Prospero, J. M. & Stowe, L. L. Characterization of tropospheric aerosols over the oceans with the advanced very high resolution radiometer optical thickness operation product, NOAA. *J. Geophys. Res.* **102**, 16889–16901 (1997).
- Measures, C. I. & Vink, S. On the use of dissolved aluminium in surface waters to estimate dust deposition to the ocean. *Glob. Biogeochem. Cycles* **14**, 317–327 (2000).
- Petit, J. R. *et al.* Climate and atmospheric history of the past 420 000 years from the Vostok ice core, Antarctica. *Nature* **399**, 429–436 (1999).
- Basile, I. *et al.* Patagonian origin of glacial dust deposited in East Antarctica (Vostok and Dome C) during glacial stages 2, 4 and 6. *Earth Planet. Sci. Lett.* **146**, 573–589 (1997).
- Diekmann, B. *et al.* Terrigenous sediment supply in the Scotia Sea (Southern Ocean): response to late Quaternary ice dynamics in Patagonia and on the Antarctic Peninsula. *Paleogeogr. Paleoclimatol. Paleoecol.* **162**, 357–388 (2000).
- Bareille, G., Grousset, F. E. & Labracherie, M. Origin of detrital fluxes in the southeast Indian Ocean during the last climatic cycles. *Paleoceanography* **9**, 799–819 (1994).
- Kumar, N. *et al.* Increased biological productivity and export production in the glacial Southern Ocean. *Nature* **378**, 675–680 (1995).
- Kidd, R. B. *et al.* King's Trough Flank: geological and geophysical investigations of its suitability for high-level radioactive waste disposal. *Inst. Oceanogr. Sci. Rep.* **166**, 1–98 (1983).
- Reiff, J., Forbes, G., Spiessma, F. & Reyniers, J. African dust reaching northwestern Europe: A case study to verify trajectory calculations. *J. Clim. Appl. Meteorol.* **25**, 1543–1567 (1986).
- Jickells, T. D. & Spokes, L. J. in *The Biogeochemistry of Iron in Seawater* (eds Turner, G. & Hunter, K. A.) (Wiley, New York, in the press).
- Martinson, D. G. *et al.* Age dating the orbital theory of the Ice Ages: Development of a high-resolution 0 to 300,000 years chronostratigraphy. *Quat. Res.* **27**, 1–29 (1987).
- Shackleton, N. J. The 100,000 year ice-age cycle identified and found to lag temperature, carbon dioxide and orbital eccentricity. *Science* **289**, 1897–1902 (2000).
- Robinson, S. G. The late Pleistocene palaeoclimatic record of North Atlantic deep-sea sediments revealed by mineral magnetic measurements. *Phys. Earth Planet. Inter.* **42**, 22–47 (1986).

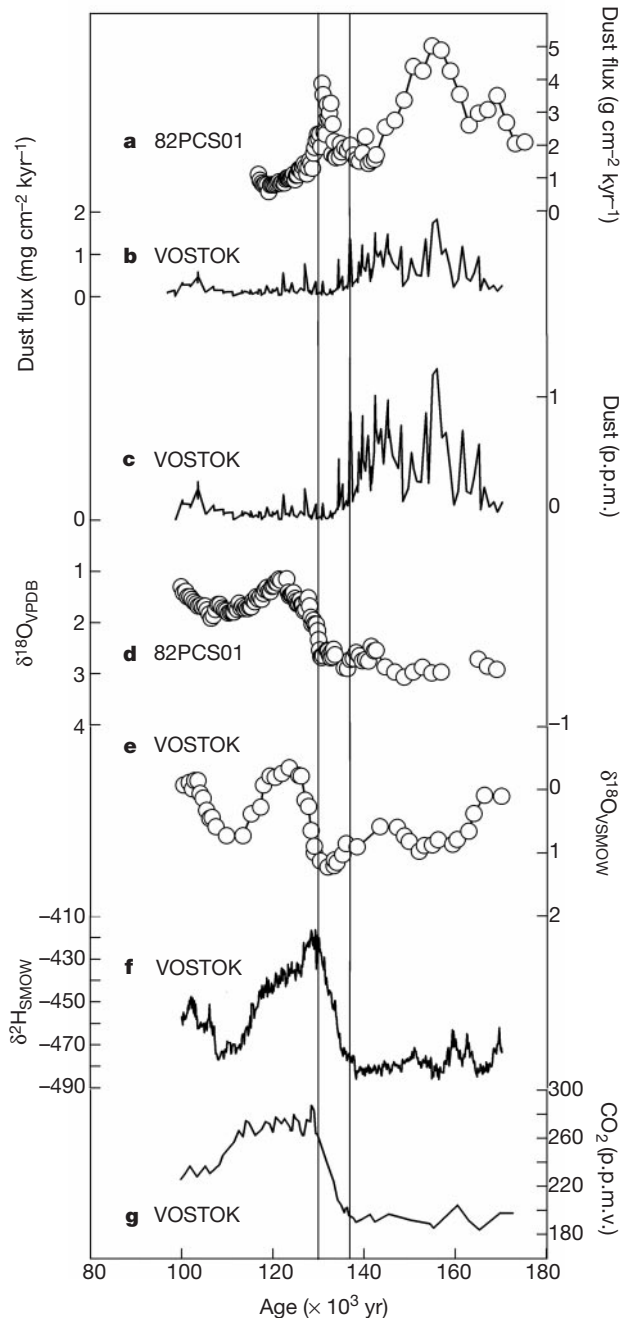


Figure 3 Dust, isotopic and CO₂ records from the North Atlantic and the Vostok ice core. **a**, Dust flux values for core 82PCS01; **b**, dust flux values for the Vostok ice core (ref. 13; J. R. Petit, personal communication); **c**, dust concentration in the Vostok ice core¹³; **d**, oxygen isotope data for piston core 82PCS01; **e**, oxygen isotope data (in air) in the Vostok ice core¹³; **f**, deuterium data for the Vostok ice core¹³; **g**, CO₂ data for the Vostok ice core¹³. The vertical lines at 137 kyr and 130 kyr mark the onset of the CO₂ rise recorded at Vostok (137 kyr) and the Termination II boundary (130 kyr).

24. Maher, B. A. & Thompson, R. in *Quaternary Climates, Environments and Magnetism* (eds Maher, B. A. & Thompson, R.) 1–48 (Cambridge Univ. Press, Cambridge, UK, 1999).
25. Chapman, M. R. & Shackleton, N. J. Millennial-scale fluctuations in North Atlantic heat flux during the last 150,000 years. *Earth Planet. Sci. Lett.* **159**, 57–70 (1998).
26. Hovan, S. A., Rea, D. K. & Pisias, N. G. Late Pleistocene continental climate and oceanic variability recorded in Northwest Pacific sediments. *Paleoceanography* **6**, 349–370 (1991).
27. DeMenocal, P. B., Ruddiman, W. F. & Pokras, E. M. Influences of high-latitude and low-latitude processes on African terrestrial climate—Pleistocene eolian records from Equatorial Atlantic-Ocean drilling Program Site-663. *Paleoceanography* **8**, 209–242 (1993).

Acknowledgements

We are grateful to Southampton Oceanography Centre for allowing sampling of Core 82PCS01. We also thank K. Miller for picking the foraminifera, and S. Dennis and G. McIntosh for performing oxygen isotope and magnetic analyses, respectively.

Correspondence and requests for materials should be addressed to B.M.
(e-mail: b.maher@lancaster.ac.uk).

Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer

Andrew M. Freed* & Jian Lin†

* Department of Terrestrial Magnetism, Carnegie Institution of Washington, Washington DC 20015, USA

† Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543 USA

Stress changes in the crust due to an earthquake can hasten the failure of neighbouring faults and induce earthquake sequences in some cases^{1–5}. The 1999 Hector Mine earthquake in southern California (magnitude 7.1) occurred only 20 km from, and 7 years after, the 1992 Landers earthquake (magnitude 7.3). This suggests that the Hector Mine earthquake was triggered in some fashion by the earlier event. But uncertainties in the slip distribution and rock friction properties associated with the Landers earthquake have led to widely varying estimates of both the magnitude and sign of the resulting stress change that would be induced at the location of the Hector Mine hypocentre—with estimates varying from –1.4 bar (ref. 6) to +0.5 bar (ref. 7). More importantly, coseismic stress changes alone cannot satisfactorily explain the delay of 7 years between the two events. Here we present the results of a three-dimensional viscoelastic model that simulates stress transfer from the ductile lower crust and upper mantle to the brittle upper crust in the 7 years following the Landers earthquake. Using viscoelastic parameters that can reproduce the observed horizontal surface deformation following the Landers earthquake, our calculations suggest that lower-crustal or upper-mantle flow can lead to postseismic stress increases of up to 1–2 bar at the location of the Hector Mine hypocentre during this time period, contributing to the eventual occurrence of the 1999 Hector Mine earthquake. These results attest to the importance of considering viscoelastic processes in the assessment of seismic hazard^{8–11}.

Earthquake triggering relationships have typically been quantified by changes in Coulomb stress occurring coseismically, $\Delta\sigma_f = \Delta\tau + \mu'\Delta\sigma_n$, where $\Delta\tau$ is the change in shear stress in the slip direction of the receiver fault, $\Delta\sigma_n$ is the change in normal stress (tension positive), and μ' is the apparent coefficient of friction incorporating the influence of pore pressure. Numerous studies have shown good correlations between calculated positive coseismic stress changes and the locations of aftershocks^{1,2,12–15} as well as

triggering of moderate to large earthquakes^{2–5,16,17}. For example, the 1933 Long Beach and 1952 Kern County earthquakes were calculated to have increased coseismic stress in the region of the 1971 San Fernando earthquake, which in turn may have increased coseismic stress at the sites of the 1987 Whittier Narrows and 1994 Northridge events¹⁶.

A limitation of the coseismic stress models is that they only consider elastic responses to fault slip, and thus cannot account for delay times in the triggering process. One possible source for postseismic stress changes is viscous relaxation. Previous analyses have shown that viscous flow in the lower crust or upper mantle after a moderate or major earthquake can lead to significant increases in stress and strain in the seismogenic upper crust, causing it eventually to become the main layer storing strains associated with the original rupture^{8–11}. Geodetic studies have revealed rapid regional-scale changes in surface deformation following the Landers earthquake^{18–22}. Though a host of processes—including fault-zone collapse¹⁹, afterslip^{18,20} and poroelastic rebound²³—may influence postseismic surface deformation, viscous flow in either the lower crust^{24,25} or (predominantly) the upper mantle²⁶ appears to best explain the observed horizontal surface deformation following the Landers earthquake. Although the viscous-upper-mantle model does a better job of explaining the observed vertical surface deformation²⁶, the viscous-lower-crust model is more consistent with petrological arguments that a felsic lower crust should be closer to its melting point and therefore weaker than a mafic upper mantle²⁷. We therefore consider both rheologies in our calculations of postseismic stress changes.

The June 1992 Landers earthquake, moment magnitude $M_w = 7.3$, was part of an earthquake sequence that began with the April 1992 $M_w = 6.1$ Joshua Tree preshock and continued with the $M_w = 6.2$ Big Bear aftershock only a few hours after the Landers rupture²⁸ (Fig. 1). To investigate the influence of viscous dissipation following the Landers sequence, we developed a three-dimensional, viscoelastic, finite-element model of a portion of the southern California lithosphere (Fig. 2). Our model considers fault slip associated with the Landers sequence in accordance with slip distributions inferred by the combined inversion of strong motion, teleseismic motion and geodetic data²⁹. Previous coseismic analyses of the Landers earthquake approximated the complex rupture surface by mapping slip distributions to three or four linear fault segments^{1,2,6,7}. Changes in Coulomb stress are, however, sensitive to kinks and jumps in the slip distribution. This is especially true for the Hector Mine hypocentre that is located only 20 km from the region of maximum slip (5–6 m) during the Landers earthquake, where rupture jumped from the Emerson to the Homestead Valley segments²⁸ (Fig. 1b). To achieve the best possible accuracy of the calculated coseismic stress changes, we therefore mapped the inferred Landers slip distribution onto fault geometry dictated by observed surface breaks²⁸ (Fig. 2c).

The reasonableness of our Landers slip model is demonstrated by the good correspondence between the calculated coseismic Coulomb stress increases and the observed aftershocks (Fig. 1). Specifically, the model predicts the location of aftershock clusters, which are especially diagnostic of Landers stress changes as opposed to more randomly distributed normal background seismicity. Of particular interest in Fig. 1b is the region of calculated positive coseismic stress change that extends to the northeast of Landers, encompassing a cluster of aftershocks that occurred near the subsequent Hector Mine epicentre. The occurrence of this aftershock cluster implies crustal weakness near the subsequent Hector Mine rupture site, suggesting that even though coseismic stress changes were not sufficient to trigger the Hector Mine earthquake, the triggering threshold was probably close. Reasonable agreement between our Coulomb stress results and aftershock locations is achieved by assuming an apparent friction coefficient in the range $0.2 < \mu' < 0.6$, with $\mu' = 0.4$ for the example shown in Fig. 1. Our

co-seismic model is also in reasonable agreement (within 10%) with the magnitude and orientation of coseismic horizontal displacement vectors as determined by GPS (Global Positioning System) and trilateration measurements²⁹.

A more direct calculation of the influence of the Landers earthquake on the triggering of the subsequent Hector Mine rupture is to

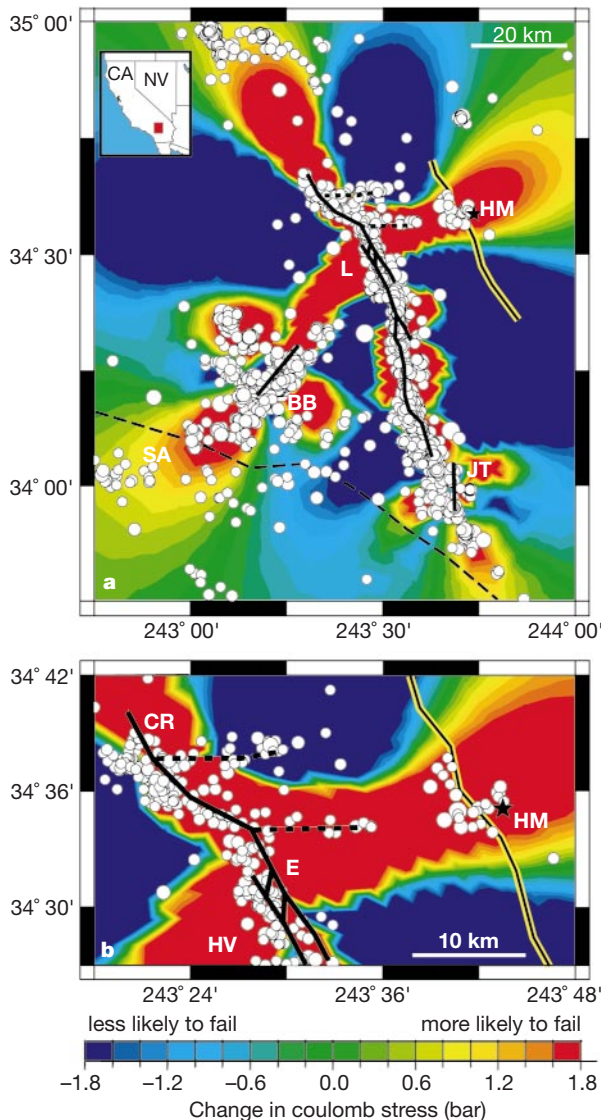


Figure 1 Correspondence between observed aftershocks and calculated coseismic changes in Coulomb stress associated with the 1992 Landers earthquake sequence.

a, Surface rupture geometry of the 1992 Landers (L), Big Bear (BB) and Joshua Tree (JT) earthquakes (solid black lines) as well as the 1999 Hector Mine (HM) earthquake (black on yellow line) and the San Andreas (SA) fault (dashed line). Hector Mine epicentre is shown by a star. The study region is in southern California (red box in inset). CA, California; NV, Nevada. **b**, Details of the northern portion of the Landers rupture region showing the Camp Rock (CR), Emerson (E), and Homestead Valley (HV) segments. Inferred buried east–west rupture surfaces²⁸ are shown as dashed lines. Seismicity ($M_w > 3$) shown (white dots) is for a period from the Landers earthquake (28 June 1992) to just before the Hector Mine earthquake (16 October 1999). Background shows the calculated coseismic Coulomb stress changes based on slip of the 1992 Landers earthquake sequence (see text). Regions of positive Coulomb stress changes are shown in red, negative changes are in blue. Receiver fault orientations were calculated considering both the coseismic stresses and a 100-bar regional compressional stress field with its maximum compression axis at $N7E^{1,2}$. The apparent coefficient of friction is $\mu' = 0.4$. All earthquakes are right-lateral strike slip except for the Big Bear earthquake that is left lateral.

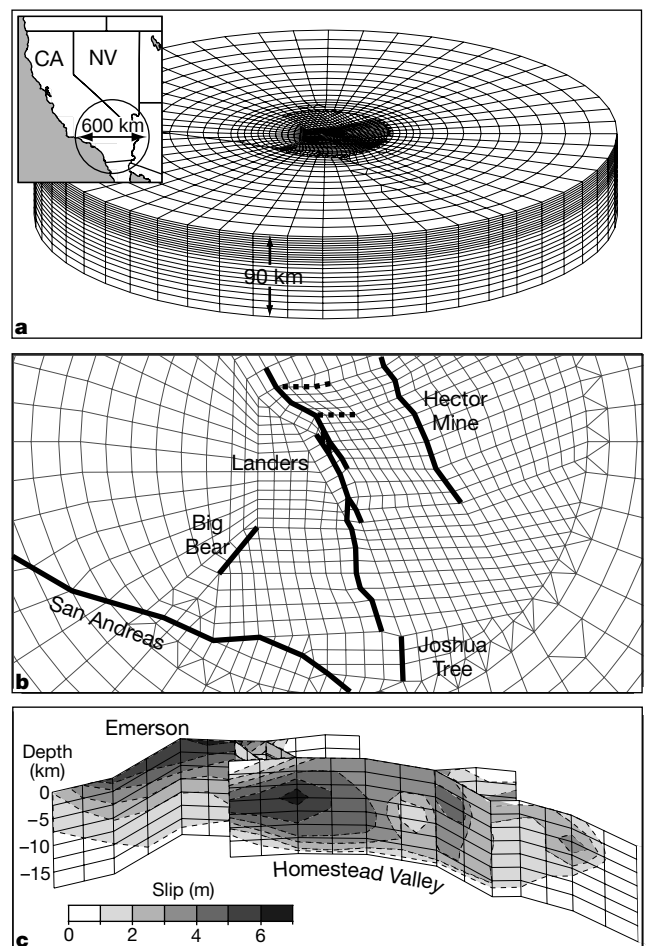


Figure 2 Viscoelastic finite-element model used to calculate coseismic and postseismic Coulomb stress changes associated with the 1992 Landers earthquake sequence. The model (**a**) comprises a cylindrical region of the southern California lithosphere with a diameter of 600 km and a depth of 90 km, centred at the Landers epicentre (inset). The cylindrical shape of the model efficiently minimizes the influence of the distant boundary conditions, which are assumed to be fixed in displacement on the edge and bottom. We used the finite-element code I-deas (see <http://www.sdrc.com>), a widely used engineering code that we have adapted for use in geodynamical problems. A close-up of the model centre detailing fault geometry is shown in **b**. Modelled slip distribution²⁹ for the Emerson and Homestead Valley segments of the Landers rupture is shown in **c**. Fault ruptures are modelled explicitly by defining slip between coincident nodes on the predefined rupture surfaces. The crust and mantle are modelled with elastic Young's moduli of 8×10^{10} and 1×10^{11} Pa, respectively, and a Poisson's ratio of 0.25. The model calculates the evolution of Coulomb stress changes associated with the relaxation of viscous lower crust and/or upper mantle. We examine two end-member cases. The first case considers viscous flow in a lower crust layer²⁵ that lies in the depth range of 18 to 28 km with a viscosity of 5×10^{18} Pa s. The second case considers predominantly viscous flow in the upper mantle²⁶, for which the lower crust has a viscosity of 1.6×10^{19} Pa s, the top portion of the upper mantle (28–50 km depth) has a viscosity of 8×10^{18} Pa s, and the lower portion of the upper mantle (50–90 km depth) has a viscosity of 5×10^{18} Pa s. Our calculations show that these sets of parameters can satisfactorily reproduce the observed horizontal surface deformation along the USGS Emerson geodetic line²⁰, with the discrepancies between models and observation $\leq 10\%$. Our model of viscous lower-crustal flow has a viscosity 5 times that suggested in ref. 25, while for our model of viscous upper-mantle flow, the mantle viscosity below 50 km is twice that suggested in ref. 26. These differences are probably due to differences in the modelled Landers slip distribution, assumed elastic properties, and our fixed boundary at a depth of 90 km. Extending the bottom boundary to a depth of 200 km changes our elastic stress results by less than 0.03 bar.

constrain the receiver faults in the Coulomb modelling to be right-lateral and to be parallel to the observed Hector Mine rupture surface (N20W). The resulting coseismic Coulomb stress changes with respect to this orientation are shown in Fig. 3a, both for the top ground surface and for a cross-sectional view of the model along the Hector Mine rupture surface. These results show that the Hector Mine hypocentre is located near a narrow strip of slightly positive Coulomb stress changes bounded by regions of negative stress changes to the north and south. However, slight changes in the assumed slip distribution and rupture geometry of the Landers earthquake can shift this region of slightly positive stress changes several kilometres to the north or south, resulting in significantly different predictions of positive or negative coseismic stress changes at the Hector Mine hypocentre. This is a major reason for the lack of a consensus^{6,7} on the sign and magnitude of the coseismic stress changes at the Hector Mine hypocentre. We calculate the coseismic Coulomb stress changes at the point of the Hector Mine hypocentre to be from -0.2 bar for a relatively high apparent friction ($\mu' = 0.8$) to -0.8 bar for a relatively low apparent friction ($\mu' = 0.2$) (see Fig. 4a for year 1992). Figure 3a displays the case of high apparent friction, as high friction along the Hector Mine rupture zone was inferred from a previous study⁷.

Of particular interest to this study is the large lobe of positive coseismic Coulomb stress changes induced in the lower crust and upper mantle by the Landers earthquake (Fig. 3a). This deep region of Coulomb stress increase is primarily driven by significant coseismic slip (5–6 m) between 5 and 12 km depth on the Homestead Valley segment of the Landers rupture zone (Fig. 2c). If either the viscous lower crust or upper mantle is unable to sustain this load following the Landers earthquake, this initial lobe of coseismic Coulomb stress increase must migrate into the upper crust as the lower crust or upper mantle relaxes. For the case of viscous lower-crustal flow, Coulomb stress at the Hector Mine hypocentre is calculated to increase by about 1 bar during the time interval from 1992 to 1999 (Fig. 3b and c). The magnitude of this increase is not very sensitive to the assumed coefficient of friction (Fig. 4a). For the case of viscous upper-mantle flow, Coulomb stress at the Hector

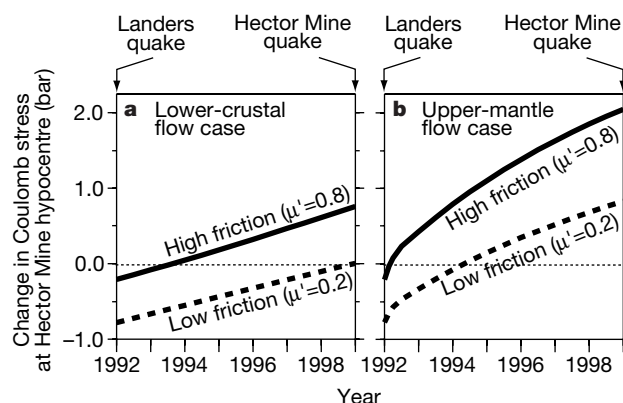


Figure 4 Calculated Coulomb stress changes at the Hector Mine hypocentre as a function of year and assumed friction. **a**, For a model that assumes viscous flow occurs only in the lower crust. **b**, For a model that assumes viscous flow occurs predominantly in the upper mantle. The calculated positive postseismic Coulomb stress changes suggest that the Hector Mine hypocentre was brought closer to failure by post-Landers relaxation processes regardless of the apparent friction or whether viscous flow occurs predominantly in the lower crust or upper mantle.

Mine hypocentre is calculated to increase by more than 2 bar in the 1992 to 1999 time interval (Fig. 3d and e). This result is also not very sensitive to the assumed friction (Fig. 4b). The calculated postseismic stress increase at the Hector Mine hypocentre is greater for the viscous upper mantle flow case because the initial reservoir of positive coseismic stress changes was greater in the upper mantle than in the lower crust (Fig. 3a).

In addition to explaining the delayed triggering of the Hector Mine earthquake, postseismic stress changes following the Landers earthquake may also have influenced the spatial extent of the Hector Mine rupture zone. Relaxation of the viscous lower crust (Fig. 3c) and upper mantle (Fig. 3e) is calculated to cause postseismic stress increase over two-thirds and the full length, respectively, of the

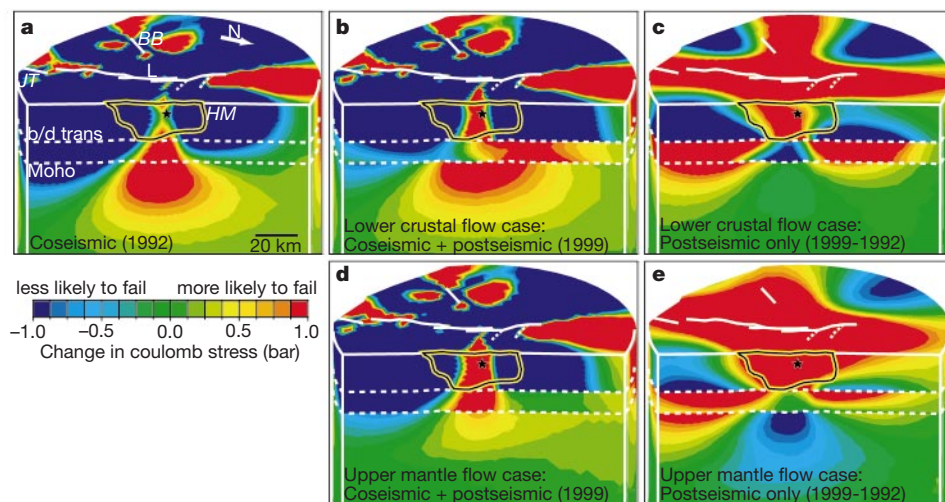


Figure 3 Calculated coseismic and postseismic changes in Coulomb stress associated with the 1992 Landers earthquake sequence. **a**, Calculated coseismic Coulomb stress changes shown both for the top ground surface and for a cross-sectional view of the model along the Hector Mine rupture surface, which is defined by regions of at least 1 m of slip³⁰ (surface encompassed by the black on yellow line). The Hector Mine hypocentre is shown by a black star. The Joshua Tree (JT), Landers (L) and Big Bear (BB) rupture surfaces are shown as white lines on the top ground surface. The lower crust lies between the brittle/ductile transition (b/d trans) at 18 km depth and the Moho at 28 km depth. The receiver faults are assumed to have the same orientation as the Hector Mine rupture zone

(N20W). All calculations shown assume apparent friction $\mu' = 0.8$. **b**, Calculated combined coseismic and 7 years of postseismic Coulomb stress changes if viscous flow occurs predominantly in the lower crust. We note that whereas stress tensors and the displacement field are continuous throughout the entire model in these calculations, the resultant Coulomb stress changes can be very sharp across the brittle/ductile transition and Moho owing to discontinuous viscoelastic properties. **c**, As **b** but for postseismic Coulomb stress changes only. **d**, Calculated combined coseismic and 7 years of postseismic Coulomb stress changes caused by predominantly viscous flow of the upper mantle. **e**, As **d** but for postseismic Coulomb stress changes only.

eventual Hector Mine rupture surface. Meanwhile, for both cases, the net (coseismic plus postseismic) changes in Coulomb stress remained negative at and beyond the distal ends of the Hector Mine rupture surface (Fig. 3b and d). Thus, it is possible that, once triggered, the Hector Mine rupture propagated to the north and south until it encountered these regions of negative net stress changes. Propagation part of the way into these negative-stress regions could be explained by the dynamics of the rupture process, which itself could have increased stress levels at the propagation fronts. Alternatively, the lateral extent of the Hector Mine rupture zone may have been controlled by the mechanical heterogeneity of the crust in the region rather than by the spatial patterns of stress changes. These results attest to the importance of considering viscoelastic processes in the assessment of seismic hazard and its migratory pattern^{8–11}. □

Received 18 September 2000; accepted 8 March 2001.

- Stein, R. S., King, G. C. P. & Lin, J. Change in failure stress on the southern San Andreas fault system caused by the 1992 magnitude = 7.4 Landers earthquake. *Science* **258**, 1328–1332 (1992).
- King, G. C. P., Stein, R. S. & Lin, J. Static stress changes and the triggering of earthquakes. *Bull. Seismol. Soc. Am.* **84**, 935–953 (1994).
- Deng, J. & Sykes, L. R. Triggering of 1812 Santa Barbara earthquake by a great San Andreas shock: Implications for future seismic hazards in southern California. *Geophys. Res. Lett.* **23**, 1155–1158 (1996).
- Harris, R. A., Simpson, R. W. & Reasenberg, P. A. Influence of static stress changes on earthquake locations in southern California. *Nature* **375**, 221–224 (1995).
- Stein, R. S. The role of stress transfer in earthquake occurrence. *Nature* **402**, 605–609 (1999).
- US Geological Survey, Southern California Earthquake Center, and California Division of Mines and Geology. Preliminary report on the 16 October 1999 M 7.1 Hector Mine, California, earthquake. *Seismol. Res. Lett.* **71**, 11–23 (2000).
- Parsons, T. & Dreger, D. S. Static-stress impact of the 1992 Landers earthquake sequence on nucleation and slip at the site of the 1999 M = 7.1 Hector Mine earthquake, southern California. *Geophys. Res. Lett.* **27**, 1949–1952 (2000).
- Pollitz, F. F. & Sacks, I. S. Consequences of stress changes following the 1891 Nobi earthquake, Japan. *Bull. Seismol. Soc. Am.* **85**, 796–807 (1995).
- Pollitz, F. F. & Sacks, I. S. The 1995 Kobe, Japan, earthquake: A long-delayed aftershock of the offshore 1944 Tonankai and 1946 Nankaido earthquakes. *Bull. Seismol. Soc. Am.* **87**, 1–10 (1997).
- Freed, A. M. & Lin, J. Time-dependent changes in failure stress following thrust earthquakes. *J. Geophys. Res.* **103**, 24393–24409 (1998).
- Deng, J., Hudnut, K., Gurnis, M. & Hauksson, E. Stress loading from viscous flow in the lower crust and triggering of aftershocks following the 1994 Northridge, California, earthquake. *Geophys. Res. Lett.* **26**, 3209–3212 (1999).
- Stein, R. S. & Lisowski, M. The 1979 Homestead Valley earthquake sequence, California: Control of aftershocks and postseismic deformation. *J. Geophys. Res.* **88**, 6477–6490 (1983).
- Oppenheimer, D. H., Reasonberg, P. A. & Simpson, R. W. Fault plane solutions for the 1984 Morgan Hill, California, earthquake sequence: Evidence for the state of stress on the Calaveras fault. *J. Geophys. Res.* **93**, 9007–9026 (1988).
- Reasenberg, P. A. & Simpson, R. W. Response of regional seismicity to the static stress change produced by the Loma Prieta earthquake. *Science* **255**, 1687–1690 (1992).
- Toda, S., Stein, R. S., Reasonberg, P. A. & Dieterich, J. H. Stress transferred by the $M_w = 6.5$ Kobe, Japan, shock: Effect on aftershocks and future earthquake probabilities. *J. Geophys. Res.* **103**, 24543–24565 (1998).
- Stein, R. S., King, G. C. P. & Lin, J. Stress triggering of the 1994 M = 6.7 Northridge, California, earthquake by its predecessors. *Science* **265**, 1432–1435 (1994).
- Jaume, S. C. & Sykes, L. R. Evolution of moderate seismicity in the San Francisco Bay region, 1850 to 1993: seismicity changes related to the occurrence of large and great earthquakes. *J. Geophys. Res.* **101**, 765–789 (1996).
- Shen, Z. *et al.* Postseismic deformation following the Landers earthquake, California, 28 June 1992. *Bull. Seismol. Soc. Am.* **84**, 780–791 (1994).
- Massonnet, D., Thatcher, W. & Vadon, H. Detection of postseismic fault-zone collapse following the Landers earthquake. *Nature* **382**, 612–616 (1996).
- Savage, J. C. & Svarc, J. L. Postseismic deformation associated with the 1992 $M_w = 7.3$ Landers earthquake, Southern California. *J. Geophys. Res.* **102**, 7565–7577 (1997).
- Jackson, D. D., Shen, Z. K., Potter, D., Ge, X. B. & Sung, L. Y. Geoscience: Southern California deformation. *Science* **277**, 1621–1622 (1997).
- Bock, Y. *et al.* Southern California permanent GPS geodetic array: Continuous measurements of regional crustal deformation between the 1992 Landers and 1994 Northridge earthquakes. *J. Geophys. Res.* **102**, 18013–18033 (1997).
- Peltzer, G., Rosen, P., Rogez, F. & Hudnut, K. Poro-elastic rebound along the Landers 1992 earthquake surface rupture. *J. Geophys. Res.* **103**, 30131–30145 (1998).
- Ivins, E. R. Transient creep of a composite lower crust, 2. A polymineralic basis for rapidly evolving postseismic deformation modes. *J. Geophys. Res.* **101**, 28005–28028 (1996).
- Deng, J., Gurnis, M., Kanamori, H. & Hauksson, E. Viscoelastic flow in the lower crust after the 1992 Landers, California, earthquake. *Science* **282**, 1689–1692 (1998).
- Pollitz, F. F., Peltzer, G. & Bürgmann, R. Mobility of continental mantle: Evidence from postseismic geodetic observations following the 1992 Landers earthquake. *J. Geophys. Res.* **105**, 8035–8054 (2000).
- Kirby, S. H. & Kronenberg, A. K. Rheology of the lithosphere: Selected topics. *Rev. Geophys.* **25**, 1219–1244 (1987).
- Hauksson, E., Jones, L. M., Hutton, K. & Eberhart-Phillips, D. The 1992 Landers earthquake sequence: Seismological observations. *J. Geophys. Res.* **98**, 19835–19858, (1993).

- Wald, D. J. & Heaton, T. H. Spatial and temporal distribution of slip for the 1992 Landers, California, Earthquake. *Bull. Seismol. Soc. Am.* **84**, 668–691 (1994).
- Dreger, D. & Kaverina, A. Seismic remote sensing for the earthquake source process and near source strong shaking: A case study of the October 16, 1999 Hector Mine Earthquake. *Geophys. Res. Lett.* **27**, 1941–1944 (2000).

Acknowledgements

We thank S. Sacks for discussions and S. Keiser for computer support. This research was supported by a NSF post-doctoral fellowship and the Southern California Earthquake Center.

Correspondence and requests for materials should be addressed to A.M.F. (e-mail: freed@dtm.ciw.edu).

Emperor penguins and climate change

Christophe Barbraud* & Henri Weimerskirch

Centre d'Études Biologiques de Chizé, Centre National de la Recherche Scientifique, 79360 Villiers en Bois, France

Variations in ocean–atmosphere coupling over time in the Southern Ocean^{1–3} have dominant effects on sea-ice extent and ecosystem structure^{4–6}, but the ultimate consequences of such environmental changes for large marine predators cannot be accurately predicted because of the absence of long-term data series on key demographic parameters^{7,8}. Here, we use the longest time series available on demographic parameters of an Antarctic large predator breeding on fast ice^{9,10} and relying on food resources from the Southern Ocean¹¹. We show that over the past 50 years, the population of emperor penguins (*Aptenodytes forsteri*) in Terre Adélie has declined by 50% because of a decrease in adult survival during the late 1970s. At this time there was a prolonged abnormally warm period with reduced sea-ice extent. Mortality rates increased when warm sea-surface temperatures occurred in the foraging area and when annual sea-ice extent was reduced, and were higher for males than for females. In contrast with survival, emperor penguins hatched fewer eggs when winter sea-ice was extended. These results indicate strong and contrasting effects of large-scale oceanographic processes and sea-ice extent on the demography of emperor penguins, and their potential high susceptibility to climate change.

Between 1952 and 2000, the emperor penguin colony located near Dumont d'Urville Station (66.7°S, 140.0°E) in Terre Adélie was monitored continuously, generating the longest data set available on an Antarctic marine predator. Data from the meteorological station 500 m from the colony shows that, after a period of stability in the 1960s (average temperatures –17.3 °C), winter temperatures began to vary extensively and were high throughout the 1970s until the early 1980s (average –14.7 °C); they then decreased but remained variable until the present time (average –16.6 °C, $t = 3.2$, $P = 0.004$ (Student's t -test; Fig. 1a). No trend was detectable for the summer temperatures. The breeding population of emperor penguins was stable until the mid-1970s, but declined abruptly by 50% in the late 1970s and has stabilized since (Fig. 1b).

We used capture–mark–recapture data spanning 1969–1989 from ringed breeding emperor penguins to estimate adult annual survival rates (see Methods). Females survived better than males but the survival curves for males and females were parallel (additive time-dependence, see Methods; Fig. 2a). Survival was particularly

* Present address: Tour du Valat, Le Sambuc, 13200 Arles, France.

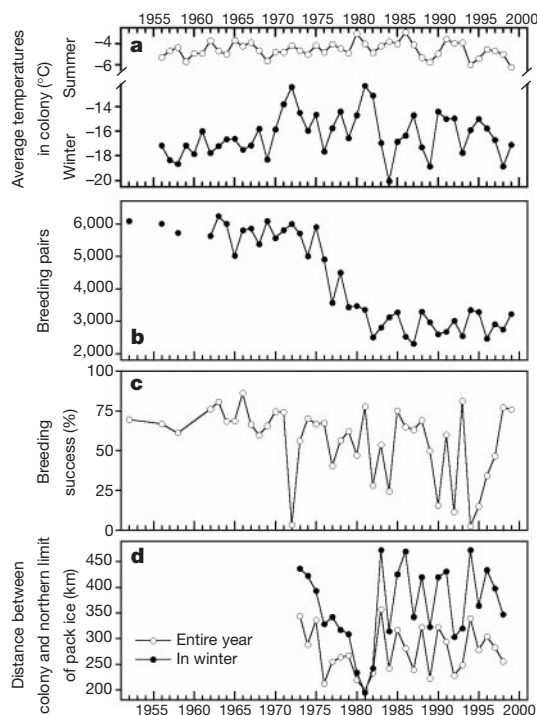


Figure 1 Climate and population changes. **a**, Summer and winter average temperatures recorded at Dumont D'Urville meteorological station (1956–1999). **b**, Number of breeding pairs of emperor penguins at Pointe Géologie archipelago, Terre Adélie (1952–1999). **c**, Changes over time in breeding success (number of chicks fledged divided by number of eggs laid). **d**, Changes over time (1973–1999) in the average distance between the emperor penguin colony and the northern extent of pack ice.

low in 1972, in 1976–1980 and in 1985 (Fig. 2a). The prolonged period of low survival during the late 1970s corresponds to the period of rapid decline of the breeding population (Fig. 2a). Interannual per cent changes in the breeding population size were related to annual survival ($P < 0.001$; Fig. 2b) and explained 64% of the variation in survival. The decline in population during the second half of the 1970s was thus caused by this unusually long period of low adult survival. The growth rate of populations of long-lived species is mainly sensitive to changes in adult survival¹². High emigration from this colony to others is unlikely as the nearest colony is 1,000 km away and penguins, like all seabirds, are faithful to their breeding site once they have started to reproduce¹³.

Among the three covariates examined (annual average sea surface temperatures (SST), northward sea-ice extent and atmospheric sea-level pressure; see Methods), SST accounted for most (89.8%) of the yearly variation in survival ($P = 0.0012$). Emperor penguins survived less when SSTs were higher ($P < 0.001$) (Fig. 2c). Sea-ice extent anomalies explained 86.9% of the variation in survival ($P = 0.0022$) and penguins survived better when sea-ice extent was greater ($P < 0.001$). Sea-level pressure did not explain a significant part of the yearly variation in survival ($P = 0.48$). To our knowledge, this is the first time that the consequences of changes in major oceanic parameters on the dynamics of an Antarctic large predator have been identified, and particularly that the proximate and ultimate factors affecting the dynamics of the population have been documented.

Because variables are interlinked, variability is generally assumed to cascade throughout the environment and biological systems. Adult emperor penguins forage only at sea and mainly feed on Antarctic krill (*Euphausia superba*), fish (mainly *Pleuragramma antarcticum*) and squid, in proportions that vary as a function of locality and time^{11,14,15}. Prey availability is influenced by physical parameters such as sea-ice extent or sea-surface temperature (or

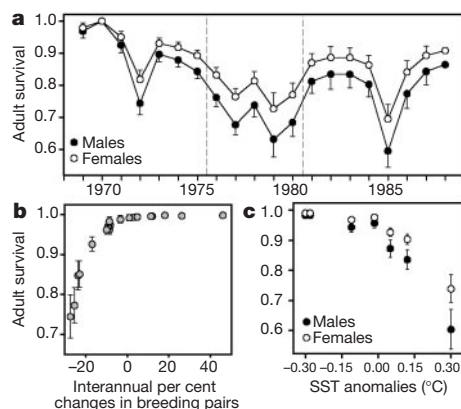


Figure 2 Adult survival of emperor penguins. **a**, Annual survival of male and female emperor penguins (mean $\pm$ 1 s.e., 1969–1988). Dashed vertical lines encompass the period of abrupt decline of the population (see Fig. 1b). The modelling of survival probabilities indicated that adult survival differed between sexes (P -likelihood ratio test (LRT) = 0.065) and varied among years (P -LRT < 0.001). Survival estimates are from model ($\Phi_{a2+s+t, p_{t,m}}$) where male and female survival are additive and time-dependent (model with lowest Akaike's Information Criterion (AIC) for the 1969–1989 data set; see Methods). **b**, Survival estimates of emperor penguins (males and females combined) from the covariate model investigating the relationship between survival and inter-annual per cent change in the breeding population size. Survival estimates are from model ($\Phi_{a2+ \text{population change}, p_{t,m}}$). Population change accounted for a large part of the annual variation in survival ($\chi^2_{17} = 18.25$, $P = 0.37$). The slope of the logit-linear relationship of survival Φ with population change (logit $\Phi = a + b \times \text{population change}^{26}$) was 0.42 ± 0.12 (lower

and upper 95% confidence limits were 0.19 and 0.66, respectively), indicating lower survival when the population decreased. Error bars, $\pm$ 1 s.e. **c**, Survival estimates of male and female emperor penguins plotted against sea-surface temperature anomalies for the period 1982–1988, when both parameters were available. Survival estimates are from model ($\Phi_{s+SST, p_{t,m}}$) with the lowest AIC value. The AIC-differences between this model and models ($\Phi_{s+t, p_{t,m}}$) and ($\Phi_{s, p_{t,m}}$) were, respectively, 4.48 and 50.51, and SST accounted for a large part of the annual variation in survival ($\chi^2_5 = 5.942$, $P = 0.312$). The slope of the logit-linear relationship of survival Φ with SST was -6.02 ± 1.01 (lower and upper 95% confidence limits -7.99 and -4.05 , respectively), indicating lower survival with higher SST. Errors bars, $\pm$ 1 s.e. The slope of the relationship of survival Φ with sea-ice extent was 0.38 ± 0.10 (lower and upper 95% confidence limits 0.18 and 0.58, respectively), indicating higher survival with greater sea-ice extent.

both)¹⁶. Decreased frequency of krill recruitment associated with a decreased frequency of extensive winter sea-ice may be responsible for low population sizes of krill¹⁷, and lower krill abundance is associated with areas with less winter sea-ice cover⁵. In years with high SSTs, emperor penguins probably have difficulties in finding food, which could increase mortality. It is not known when the survival of adult emperor penguins is affected by warm water events: whether during winter when males and females undertake long incubation and chick brooding fasts^{18,19}, or during the summer when they disperse to moult²⁰.

Breeding success varied extensively throughout the period, and its variability has increased progressively since the 1970s (Fig. 1c). A combination of local factors has probably contributed to the high variability in breeding success. Complete or extensive breeding failures in some years resulted from early break-out of the sea-ice holding up the colony, or from prolonged blizzards during the early chick-rearing period²¹. Overall breeding success was not related to SST anomalies or sea-ice extent, possibly because variation was the result of a combination of confounding factors related to sea-ice conditions or weather conditions. However, the proportion of eggs that hatched a chick was negatively related to the extent of pack ice in winter ($r^2 = 0.284$, $P = 0.02$), with wider pack ice resulting in lower hatching success.

One paradoxical and unexpected result of this study is that the extent of pack ice, which is a key factor in the Antarctic ecosystem⁷, has two opposite effects on the demographic parameters of emperor penguins. Sea-ice extent in winter negatively affects hatching success, by increasing the distance between the colony and feeding grounds. Conversely, annual sea-ice extent positively affects adult survival by increasing food availability. Therefore there exists in emperor penguins a trade-off between the advantages and disadvantages of extensive pack ice. In population terms, the trophic advantage of extensive pack ice, by favouring higher survival and further reproduction, outmatches its physical disadvantage of reducing fecundity.

Whereas time series on emperor penguins and on meteorological conditions in the colony are available since 1952, data on the extent of sea-ice are available only from 1972 onwards, and on SST from 1982. Given the strong relationship between survival and SST anomalies related to the Antarctic Circumpolar Wave² in 1982–1989, low survival during 1976–1980 (and the 50% population decrease) was probably associated with high SSTs, but remote-sensing satellite data were not available at this time. The temperature recorded in winter in the colony, however, indicates a prolonged warm period during the 1970s (Fig. 1a) associated with a prolonged period of decreasing extent of ice¹ (Fig. 1d). This anomaly is probably connected to the abrupt decline in Antarctic sea-ice extent that is apparent from whaling records during the 1960s and 1970s^{1,3}. The exact timing of the decline is not known because of the lack of records during this period³. Our results indicate that this major event may have occurred during the 1970s rather than during the 1960s as suggested earlier. Our results also indicate that emperor penguins may be very susceptible to environmental variability and that further long-lasting coupled anomalies are likely to affect their populations. □

Methods

We determined the number of breeding pairs from the number of chicks fledged plus the number of eggs and chicks found dead. Dead chicks and eggs remain frozen on sea-ice around the colony and were collected and counted daily during the entire breeding season. The number of breeding pairs represents 80% of the overall population size.

Environmental parameters

We used three types of data: the annual average SST calculated from a combination of *in situ* and satellite-radiometer measurements obtained on a 1° grid²²; the annual average northward sea-ice extent calculated from space-borne microwave sensors on a 25-km grid²³; and the annual averages of atmospheric sea-level pressure calculated from data produced by the European Centre for Medium Range Weather Forecasts on a 1.875° grid²⁴, all spanning 1982–1989. Data were extracted for the 200-km radius (from latitude 66.7° S)

surrounding the Pointe Géologie colony, an area consistent with the foraging radius of emperor penguins^{11,19}. Physical parameters were determined relative to the applicable record length annual mean values and filtered to suppress seasonal and possible biennial signals².

Survival modelling

Adults were ringed each year between 1967 and 1980 using metal bands placed at the base of the flippers. Ringing and ring readings were made in April and May when birds arrived on the breeding grounds forming columns and during pair formation from 1968 until today⁹. Because during the first year after banding, and to a lesser extent during the second year, bands are more likely to be lost or to affect survival²⁵, we have excluded data from the first two years after banding. We estimated adult survival rate using recent developments of the Cormack–Jolly–Seber model^{26,27} implemented in the MARK software²⁸ for the period 1969–1988. This method allows the estimation of unbiased capture and survival probabilities. Parameters that may explain time-variation in survival were then examined using the capture–mark–recapture data set over the period 1982–1989, when data were simultaneously available with adult survival (before 1982, no physical data were available). The 1982–1989 data set was used to test the effects of physical parameters on survival. Details of survival modelling procedures are available as Supplementary Information.

Changes in breeding population size from year to year should be correlated with survival if variations in the latter are a principal cause of the former. To test whether the temporal variation in survival is likely to have been a significant cause of changes in population size we used a covariate model in which the interannual per cent change in the breeding population size was added to the selected model as an annual covariate²⁹ with a one-year lag, as the breeding population size of emperors was determined in May.

Received 23 October 2000; accepted 20 February 2001.

- Kukla, G. & Gavin, J. Summer ice and carbon dioxide. *Science* **214**, 497–503 (1981).
- White, W. B. & Peterson, R. G. An Antarctic circumpolar wave in surface pressure, wind, temperature and sea-ice extent. *Nature* **380**, 699–702 (1996).
- de la Mare, W. K. Abrupt mid-twentieth century decline in Antarctic sea-ice extent from whaling records. *Nature* **389**, 57–60 (1997).
- Tynan, C. T. Ecological importance of the Southern Boundary of the Antarctic Circumpolar current. *Nature* **392**, 708–716 (1998).
- Nicol, S. *et al.* Ocean circulation off east Antarctica affects ecosystem structure and sea-ice extent. *Nature* **406**, 504–507 (2000).
- Aebischer, N. J., Coulson, J. C. & Colebrook, J. M. Parallel long-term trends across four marine trophic levels and weather. *Nature* **347**, 753–755 (1990).
- Croxall, J. P. Southern Ocean environmental changes: effects on seabird, seal and whale populations. *Phil. Trans. R. Soc. Lond. B* **338**, 319–328 (1992).
- Furness, R. W. & Greenwood, J. J. D. (eds) *Birds as Monitors of Environmental Change* (Chapman and Hall, London, 1993).
- Prévost, J. *Ecologie du Manchot Empereur Aptenodytes forsteri* Grey (Expéditions Polaires Françaises, Hermann, Paris, 1961).
- LeMaho, Y. The Emperor penguin: a strategy to live and breed in the cold. *Am. Sci.* **65**, 680–693 (1977).
- Kirkwood, R. & Robertson, G. The foraging ecology of female Emperor Penguins in winter. *Ecol. Monogr.* **67**, 155–176 (1997).
- Lebreton, J. D. & Clobert, J. in *Bird Population Studies* (eds Perrins, C. M., Lebreton, J. D. & Hiron, G. J. M.) 103–125 (Oxford Univ. Press, Oxford, 1991).
- Williams, T. D. *The Penguins* (Oxford Univ. Press, Oxford, 1995).
- Offredo, C. & Ridoux, V. The diet of Emperor penguins *Aptenodytes forsteri* in Adélie land, Antarctica. *Ibis* **128**, 409–413 (1986).
- Cherel, Y. & Kooyman, G. L. Food of emperor penguins (*Aptenodytes forsteri*) in the western Ross Sea, Antarctica. *Mar. Biol.* **130**, 335–344 (1998).
- Pakhomov, E. A. & McQuaid, C. D. Distribution of surface zooplankton and seabirds across the Southern Ocean. *Polar Biol.* **16**, 271–286 (1996).
- Loeb, V. *et al.* Effects of sea-ice extent and krill or salp dominance on the Antarctic food web. *Nature* **387**, 897–900 (1997).
- Iseemann, P. Contribution à l'éthologie et à l'écologie du Manchot empereur (*Aptenodytes forsteri* Gray) à la colonie de Pointe Géologie (Terre Adélie). *Oiseau R.F.O.* **41**, 9–61 (1971).
- Ancel, A. *et al.* Foraging behaviour of emperor penguins as a resource detector in winter and summer. *Nature* **360**, 336–338 (1992).
- Kooyman, G., Hunke, E., Ackley, S., van Dam, R. & Robertson, G. Moulting of the emperor penguin: travel, location, and habitat selection. *Mar. Ecol. Prog. Ser.* **204**, 269–277 (2000).
- Jouventin, P. in *The Biology of Penguins* (ed. Stonehouse, B.) 435–446 (Macmillan, London, 1974).
- Reynolds, R. W. & Marisco, D. C. An improved real time global sea surface temperature analysis. *J. Climate* **6**, 114–119 (1993).
- Parkinson, C. L. Spatial patterns in the length of the sea ice season in the Southern Ocean, 1979–1986. *J. Geophys. Res.* **99**, 16327–16339 (1994).
- ECMWF/WCRP. *A Global Atmospheric Data Archive* (Tech. Attach. European Centre for Medium Range Weather Forecasts, Reading, 1993).
- Weimerskirch, H., Stahl, J. C. & Jouventin, P. The breeding biology and population dynamics of king penguins *Aptenodytes patagonicus* at the Crozet islands. *Ibis* **134**, 107–117 (1992).
- Nichols, J. D. Capture–recapture models: using marked animals to study population dynamics. *Bioscience* **42**, 94–102 (1991).
- Lebreton, J. D., Burnham, K. P., Clobert, J. & Anderson, D. R. Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecol. Monogr.* **62**, 67–118 (1992).
- White, G. C. & Burnham, K. P. Program MARK: survival estimation from populations of marked animals. *Bird Study* **46**, 120–138 (1999).
- Siriwardena, G. M., Baillie, S. R. & Wilson, J. D. Temporal variation in the annual survival rates of six granivorous birds with contrasting population trends. *Ibis* **141**, 621–636 (1999).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This study was supported over the past 50 years by Expéditions Polaires Françaises, Terres Australes et Antarctiques Françaises and by Institut Français pour la Recherche et la Technologie Française. We thank all the people involved in Terre Adélie in the monitoring programs of the Emperor penguin population, and Y. Cherel, G. Kooyman and C. Parkinson for comments on the manuscript. We thank Météo France DOM-TOM, G. Testa and M. Gaillot for the meteorological data of the station of Dumont D'Urville, Terre Adélie.

Correspondence and requests for materials should be addressed to C.B. (e-mail: barbraud@tour-du-valat.com).

Host recognition by the tobacco hornworm is mediated by a host plant compound

Marta L. del Campo^{*†}, Carol I. Miles^{†‡}, Frank C. Schroeder[§], Caroline Mueller^{||}, Ronald Booker[‡] & J. Alan Renwick^{||}

^{*} Department of Entomology, Cornell University, Ithaca, New York 14853, USA

[†] Department of Biological Sciences, Binghamton University, PO Box 6000 Binghamton, New York 13902-6000, USA

[‡] Department of Neurobiology and Behavior, Cornell University, Ithaca, New York 14853, USA

[§] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA

^{||} Boyce Thompson Institute at Cornell University, Ithaca, New York 14853, USA

It is generally believed that animals make decisions about the selection of mates, kin or food on the basis of pre-constructed recognition templates. These templates can be innate or acquired through experience¹. An example of an acquired template is the feeding preference exhibited by larvae of the moth, *Manduca sexta*. Naïve hatchlings will feed and grow successfully on many different plants or artificial diets, but once they have fed on a natural host they become specialist feeders^{2–6}. Here we show that the induced feeding preference of *M. sexta* involves the formation of a template to a steroidal glycoside, indioside D, that is present in solanaceous foliage. This compound is both necessary and sufficient to maintain the induced feeding preference. The induction of host plant specificity is at least partly due to a tuning of taste receptors to indioside D. The taste receptors of larvae fed on host plants show an enhanced response to indioside D as compared with other plant compounds tested.

By manipulating the diets of larvae, we examined the strength of induced feeding preferences. *Manduca sexta* larvae were reared from hatching until the early fourth instar on either solanaceous or non-solanaceous diets. The solanaceous diets consisted of foliage from potato (*Solanum tuberosum*) or tomato (*Lycopersicon esculentum*); the non-solanaceous diets were either foliage of cowpea (*Vigna sinensis*, Fabaceae), or a wheat-germ-based artificial diet⁷. After the molt to the fourth instar, larvae from these four groups were removed from their original diet, and randomly assigned to one of the two solanaceous plants, or to the non-solanaceous cowpea. Within 24 h, all larvae reared on solanaceous foliage had begun feeding on solanaceous leaves; however, only 21.5% (11/51) of the solanaceous-reared larvae had fed on the foliage of the non-solanaceous cowpea (Fig. 1a). After 96 h, the percentage of solanaceous-reared larvae feeding on cowpea had increased to 45.1% (23/51), although by this time 35.3% (18/51) of the larvae had starved to death, having never fed on cowpea (Fig. 1b). In contrast, after only 24 h, all the larvae reared on non-solanaceous diets had accepted

either the solanaceous or non-solanaceous foliage equally (Fig. 1a), and no mortality was observed among these larvae. These results show that once host plant specificity is induced, some *M. sexta* larvae will actually starve to death rather than change from a solanaceous to a non-solanaceous diet.

Because contact with solanaceous foliage seems necessary to induce a preference for the host plant, the larvae may acquire the ability to recognize a specific compound(s) in the host plant. To test this possibility, we began to isolate a compound(s) from solanaceous plants that the larvae might use as a host plant recognition cue. As an assay, we tested solanaceous-reared larvae for their acceptance of non-solanaceous cowpea leaf discs that were treated with extracts of potato leaves. The larvae clearly discriminated between treated and control discs, taking the first bite from cowpea discs treated with potato extract in more than 95% of the cases ($n = 75$, $\chi^2 > 200$, degrees of freedom (d.f.) = 1, $P < 0.001$), in a period of less than 2 h (Fig. 2a).

An active fraction was isolated from an aqueous extract of potato foliage by solvent partitioning and flash chromatography. This active fraction was subjected to further fractionation and exhaustive chromatographic separation. At each step, activity was monitored using choice assays with fraction-treated cowpea leaf discs and ten or more solanaceous-reared larvae. This approach yielded a single fraction that accounted for all the feeding stimulant activity of potato foliage. Analyses by ¹H NMR spectroscopy showed that this fraction comprised one single steroidal glycoside of more than 96% purity. A standard set of two-dimensional NMR experiments⁸ and additional mass spectroscopic analyses using negative-ion electrospray ionization established the structure (Fig. 2c) of this compound as (25R)-26-O-(β -D-glucopyranosyl)-furost-5-en-3 β ,22 ξ ,26-triol-3-O-(α -L-rhamnopyranosyl-(1 $\rightarrow$ 2)-[β -D-glucopyranosyl-(1 $\rightarrow$ 3)]- β -D-galactopyranoside) (indioside D), which was isolated previously from the solanaceous *Solanum indicum*⁹.

Whereas considerable attention has focused on the theoretical aspects of template recognition, little is known about its proximate mechanisms^{1,10,11}. Our behavioural data suggest that induced *M. sexta* larvae use indioside D to recognize their host plant. Evidence suggests that the lateral and medial sensilla styloconica on the mouthparts (Fig. 3a, b) of *M. sexta* have a key role in taste and food preferences^{12–16}. To examine the role of these taste sensilla in host recognition, we recorded their electrophysiological responses to indioside D. The sensilla recordings were carried out on fifth instar larvae that were reared on either solanaceous foliage or wheat germ diet. Each recording electrode contained either control solu-

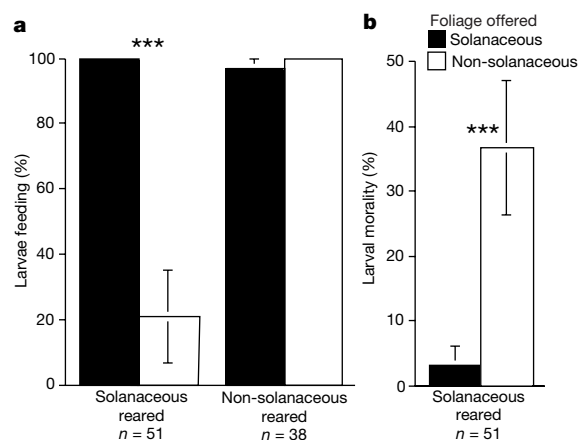


Figure 1 Host-restricted feeding behaviour is determined by dietary experience. **a**, Percentage of larvae feeding on foliage in no-choice tests, after a 24-h exposure to novel food item (Fisher's exact test). **b**, Percentage of larval mortality among solanaceous-reared larvae after a 96-h exposure to fresh foliage (Fisher's exact test). Error bars, 95% confidence interval. Triple asterisk, $P < 0.001$.

tion, comprising the KCl recording solution alone, or indioside D solution, comprising the recording solution plus $1.2 \mu\text{M}$ indioside D. The response of each sensillum to the control solution was compared with its response to the indioside D solution during the first minute after contact.

The medial sensilla showed similar increases in response to indioside D in larvae reared on either diet (paired t -tests; $n = 8$ per group; $P < 0.05$) suggesting that dietary experience had no significant effect on the responses of the medial sensilla to indioside D. In contrast, whereas the lateral sensilla of the wheat germ diet-reared larvae showed no significant difference in their responses to indioside D compared to the control solution (paired t -test, $n = 8$ per group, $P > 0.05$), the lateral sensilla of the solanaceous-reared larvae exhibited a significant increase in response to indioside D when compared with the control solution (paired t -test, $n = 8$ per group, $P < 0.01$), suggesting that dietary experience modified the responses of these sensilla. These results indicated that dietary experience might alter the responsiveness of the lateral sensilla to different plant compounds.

To test this possibility, we examined the response of the lateral sensilla of both artificial-diet- and solanaceous-reared larvae to KCl, NaCl, glucose, tomatine and indioside D. The results showed that diet could influence the response of the lateral sensilla to a range of plant compounds (Fig. 3d). Responses to the other compounds tested were either unchanged (NaCl and tomatine, t -tests, $P > 0.05$) or reduced (KCl and glucose t -tests, $P < 0.05$) as a result of experience with solanaceous foliage. In contrast, there was an increase in the responses of the lateral sensilla on solanaceous-reared larvae to indioside D (t -test, $P < 0.05$). These results indicate that the lateral sensilla may become tuned to indioside D as larvae feed on solanaceous foliage, by markedly increasing the difference between their response to the recognition cue, indioside D, and responses to other plant compounds.

We next determined whether the inputs from one or both sensilla types are required to maintain host specificity, by carrying out a series of sensillar ablations on solanaceous-reared larvae. The larvae were tested for acceptance or rejection of cowpea leaf discs before

and 24 h after the removal of all four sensilla, the bilateral removal of the medial or lateral sensilla, or the unilateral removal of both sensilla. We used larvae in which neighbouring mechanosensory hairs had been cut as controls. All control larvae ($n = 8$) and those that received unilateral sensillar ablations ($n = 10$) continued to reject cowpea. However, 100% of the larvae that had all four sensilla removed ($n = 11$), or were subjected to bilateral removal of either the medial ($n = 16$) or lateral ($n = 14$) sensilla fed readily on the non-solanaceous cowpea leaf discs. These results indicate that host-restricted feeding behaviour of solanaceous-reared larvae is dependent on sensory inputs from both the lateral and medial sensilla.

Changes in the responsiveness of taste receptors owing to dietary experience have been reported in other studies of insect feeding preferences^{17–21}. For example, in the locust, a generalist herbivore, taste receptor responses to amino acids, salts and sugars can be modified by feeding on a low protein diet²⁰. In this case, the sensory neurons exhibit increased sensitivity to a range of compounds found in common to most plants. Clearly, the strategy for a specialist insect has to be different. Ideally, the range of chemicals that could serve as potential cues should be limited to those that are unique to the host range. The more restricted the distribution of the cues, the lower the probability of mistakenly accepting a non-host plant. Thus, we predict that the distribution of indioside D and structurally related compounds will be restricted to Solanaceae.

As facultative specialists, *M. sexta* larvae can grow and complete their development on many non-host plants^{2–5,22}. Because adult females occasionally lay their eggs on non-solanaceous foliage⁶, it is essential that the larvae can feed on these plants, or they would quickly starve to death. However, larvae reared on solanaceous foliage grow and develop far more quickly than on non-solanaceous plants (M.L.d.C. and J.A.R., unpublished data). Rapid growth during the larval period might be advantageous to insects, resulting in less time spent in a life stage that is vulnerable to predators, pathogens, parasitoids or abiotic environmental hazards. Thus, it is beneficial for *M. sexta* larvae to be facultative rather than strict specialist or generalist feeders.

Here we have shown that contact with a single compound,

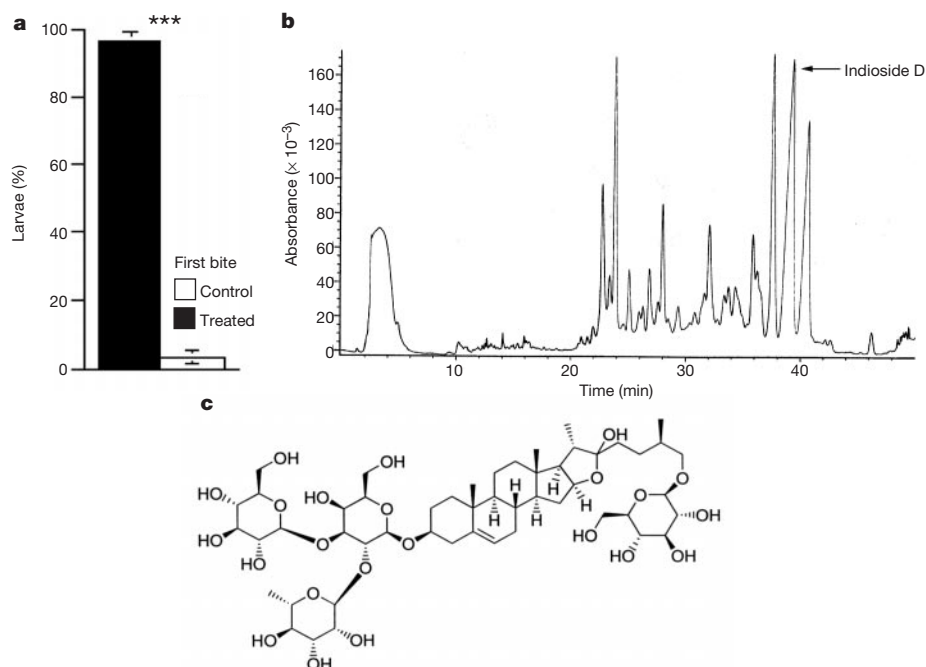


Figure 2 Indioside D is a recognition cue for host-restricted larvae. **a**, Percentage of solanaceous-reared larvae taking their first bite from cowpea disc treated with aqueous potato extract or solvent alone (control) in choice assays ($n = 75$, $\chi^2 > 200$, d.f. = 1, $P < 0.001$). Error bars, 95% confidence interval. Triple asterisk, $P < 0.001$. **b**, HPLC

chromatogram ($\lambda = 212$) of preliminary active fraction, in which Indioside D can be detected as a single peak. Other compounds present in this fraction were inactive as feeding recognition cues for solanaceous-reared *M. sexta* larvae. **c**, Chemical structure of indioside D.

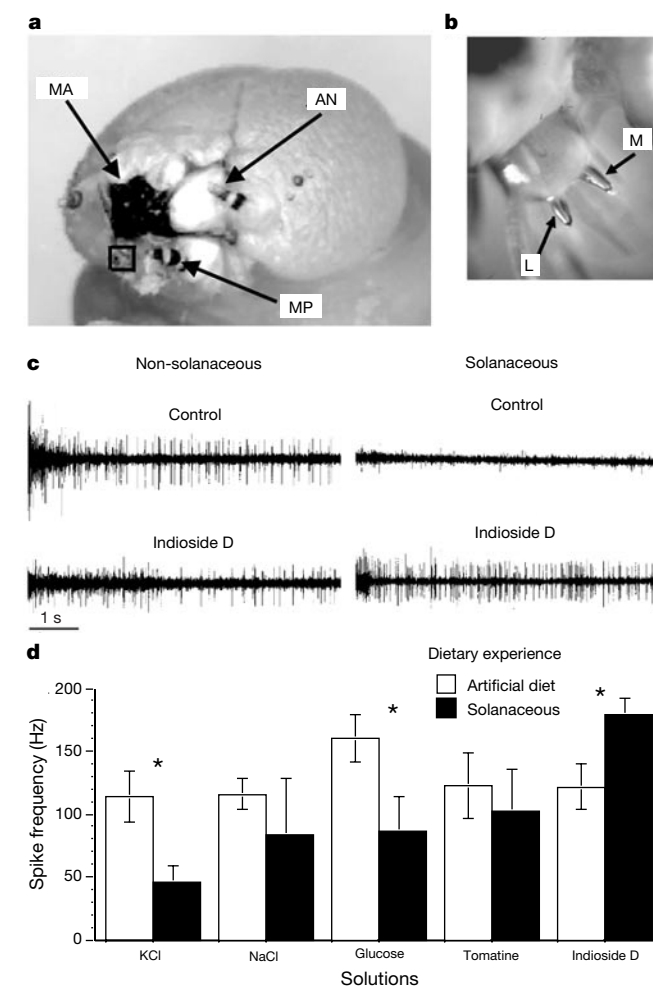


Figure 3 The lateral sensilla styloconia of host-restricted larvae are tuned to indioside D. **a**, Lateral view of head and mouthparts of *M. sexta*. AN, antenna; MA, mandible; MP, maxillary palp. **b**, Enlargement of the boxed area in **a** showing the medial (M) and lateral (L) sensilla styloconia. **c**, Electrophysiological recordings of the lateral sensilla styloconia of *M. sexta* larvae reared on non-solanaceous or solanaceous diets to control solution and indioside D solution (1.2 μ M indioside D plus control solution). **d**, Average spike frequency of response of the lateral sensilla styloconia during the first 200 ms after contact in larvae raised on artificial diet or solanaceous foliage ($n = 6$ larvae per category) to 50 mM KCl, 50 mM NaCl, 100 mM glucose, 100 μ M tomatine and 1.2 μ M indioside D. Error bars, 95% confidence intervals. Asterisk, $P < 0.05$.

indioside D, that is isolated from a host plant can maintain an established feeding specialization. Indioside D alone is both necessary and sufficient to initiate feeding bouts by host-restricted *M. sexta* larvae. We have also shown that maintaining this preference involves an alteration in the responsiveness of sensory neurons located on the mouthparts of larval *M. sexta*. Our results suggest that the host plant recognition template acquired through dietary experience is, at least in part, due to the tuning of the lateral sensilla styloconia to a chemical cue in the host. This tuning is accomplished by increasing the responsiveness of the lateral sensilla specifically to indioside D, and not to the other plant compounds that we tested. The identification of a host plant compound used as a host recognition cue, which modifies sensory responses of taste receptors and ultimately the animal's behaviour, may be a key to understanding the neural mechanisms underlying recognition. □

Methods

Choice assay

The potato aqueous extract was tested in a choice assay using 1.4-cm diameter discs cut from fresh cowpea leaves as the substrate. Four discs were presented to each larva. To each

side of every disc, we evenly applied 20 μ l of solution using a motorized micropipette. Two of the discs were treated with the extract diluted in methanol, and two were treated with methanol only as controls. Each treated disc received roughly 4.8 μ M indioside D. The four discs were pinned in a Petri dish (15-cm diameter) in an alternating fashion, and wet filter paper lined the base of the dish to provide moisture. Larvae reared on solanaceous foliage were assigned randomly and placed at the centre of individual dishes. We recorded whether a treated or control disc was bitten first, and the time taken to start feeding in each case.

Chemical extraction and isolation

Potato aqueous extract was prepared by boiling potato leaves in distilled water, cooling in an ice-water bath, and filtering. The pH was increased (> 11.0) by adding 10 M NaOH before partitioning with *n*-butanol. The butanol extract was evaporated *in vacuo* at 50 $^{\circ}$ C, and the residue was recovered in distilled water and filtered; this butanol extract was active, whereas the post-butanol aqueous extract was not. C_{18} reversed-phase flash chromatography was used to fractionate the sample further using increasing concentrations of MeOH in water. Activity was obtained in fractions eluted with 50–80% MeOH; the fractions were combined and evaporated, and the residue was recovered in water. The active fraction was passed through C_{18} Sep-Pak cartridges (Waters) with increasing concentrations of MeOH in 1% ammonia. Activity was obtained only in the fraction eluting at 60% MeOH. The active fraction was further fractionated by reversed-phase high performance liquid chromatography (HPLC; Luna reversed-phase column, C_{18} , 250 $\times$ 10.00 mm 5 μ m, Phenomenex) with a gradient of 100% water to 100% acetonitrile. Final fractionation to obtain a pure active fraction was performed in HPLC using an analytical reversed-phase C_{16} amide Discovery column (Supelco). We collected a single pure compound that accounted for all the activity.

Chemical analyses

We analysed a 1.6-mg sample of isolated material by NMR spectroscopy using a Varian UNITY+ (500-MHz proton, 126-MHz carbon) spectrometer. Spectra were recorded at 26.5 $^{\circ}$ C using methanol- D_3 and pyridine- D_3 as solvents. Phase-sensitive double-quantum filtered correlation spectroscopy (dq-COSY) spectra, phase-sensitive nuclear Overhauser effect spectroscopy (NOESY) spectra, and non-gradient heteronuclear multiple-bond correlation (HMBC) spectra were acquired using the standard pulse sequences and phase cycling⁷. From these data, we determined carbon and proton chemical-shift values, as well as most of the proton–proton coupling constants in the pentacyclic steroidal core and in the four carbohydrate units. Carbon chemical-shift data are in good agreement with published values⁸. Proton chemical-shift data and coupling constants will be published elsewhere. Mass spectroscopic analyses were performed using a Micromass Quattro I mass spectrometer operated in negative ion electrospray mode. The most abundant peaks correspond to $(M-H)^- = 1063.6$ a.m.u., $(M-C_6H_{11}O_4)^- = 917.6$ a.m.u., and $(M-C_6H_{11}O_5)^- = 901.6$ a.m.u. (M, molecular mass of the most abundant species of $C_{51}H_{84}O_{23}$).

Electrophysiological recordings

Larvae were anaesthetized by submerging them in water. We prepared them for recording by placing them in vials of water so that all spiracles remained submerged with the head exposed for recording, using a modification of described techniques²³. The head was fixed into position with soft wax so that the medial and lateral sensilla styloconia were accessible to recording electrodes. A ground electrode was inserted into the mandibular musculature through a small hole in the head capsule. Tip recordings were performed on the sensilla using glass microelectrodes filled with solutions described in the text. Signals were amplified with a high impedance amplifier (Getting Instruments). They were subsequently either stored in an instrumentation tape recorder (Vetter) and analysed on playback through a tunable active filter (Krohn-Hite) using Igor Pro (Wave Metrics), or recorded and analysed on a computer using an analog-to-digital circuit (Axon Instruments). All test solutions (50 mM KCl, 50 mM NaCl, 100 mM glucose, 100 μ M tomatine and 1.2 μ M indioside D) contained 1% methanol, 0.1% ethanol and 0.16% polyvinylpyrrolidone to increase viscosity and reduce evaporation. Glucose, tomatine and indioside D were diluted in 50 mM KCl conducting solution.

Sensillar ablations

Larvae were reared on tomato or potato foliage up to their fourth or fifth instar. They were subjected to a choice test (as described above). Only those larvae that fed exclusively on cowpea leaf discs treated with potato extract, rejecting control cowpea discs, were used for further study. Sensillar ablations were quickly performed on these larvae under CO_2 anaesthesia, using fine forceps. We placed larvae back on their former foliage to recover. They were tested for acceptance or rejection of cowpea leaf discs the next day.

Received 31 October 2000; accepted 12 March 2001.

- Sherman, P. W., Reeve, H. K. & Pfennig, D. W. In *Behavioral Ecology An Evolutionary Approach* (eds Krebs, J. R. & Davies, N. B.) 69–96 (Blackwell Science, Oxford, 1997).
- Schoonhoven, L. M. Loss of host plant specificity by *Manduca sexta* after rearing on an artificial diet. *Entomol. Exp. Appl.* **10**, 270–272 (1967).
- Jerry, T., Hanson, F. E. & Dethier, V. G. Induction of specific food preference in lepidopterous larvae. *Entomol. Exp. Appl.* **47**, 211–230 (1967).
- Yamamoto, R. T. & Fraenkel, G. S. The specificity of the tobacco hornworm, *Protoparce sexta*, to solanaceous plants. *Ann. Entomol. Soc. Am.* **53**, 503–507 (1960).
- Yamamoto, R. T. Induction of hostplant specificity in the tobacco hornworm, *Manduca sexta*. *J. Insect Physiol.* **20**, 641–650 (1974).

6. Mechaber, W. L. & Hildebrand, J. G. Novel non-solanaceous host plant record for *Manduca sexta* (Lepidoptera: Sphingidae) in the southwestern United States. *Ann. Entomol. Soc. Am.* **93**, 447–451 (2000).
7. Bell, R. A., Owens, C. D., Shapiro, M. & Tardif, J. R. in *The Gypsy Moth: Research Toward Integrated Pest Management* (eds Doane, C. C. & McManus, M. L.) 599–633 (USDA-FS Technical Bulletin 1584 (Washington DC, 1981).
8. Croasmun, W. A. & Carlson, R. M. K. (eds) *Two-Dimensional NMR spectroscopy. Applications for Chemists and Biochemists* 2nd edn (VCH, New York, 1994).
9. Yajara, S. et al. Steroidal glycosides, indiosides A–E, from *Solanum indicum*. *Phytochemistry* **43**, 1319–1323 (1996).
10. Dethier, V. G. & Crnjar, R. M. Candidate codes in the gustatory system of caterpillars. *J. Gen. Physiol.* **79**, 549–569 (1982).
11. Bernays, E. A. Selective attention and host-plant specialization. *Entomol. Exp. Appl.* **80**, 125–131 (1996).
12. Waldbauer, G. P. & Fraenkel, G. Feeding on normally rejected plants by maxillectomized larvae of the tobacco hornworm, *Protoparce sexta* (Lepidoptera, Sphingidae). *Ann. Ent. Soc. Am.* **54**, 477–485 (1961).
13. de Boer, G. Role of bilateral chemosensory input in food discrimination by *Manduca sexta*. *Entomol. Exp. Appl.* **61**, 159–168 (1991).
14. de Boer, G. Effect of diet experience on the ability of different larval chemosensory organs to mediate food discrimination by the tobacco hornworm, *Manduca sexta*. *J. Insect Physiol.* **37**, 763–769 (1991).
15. Flowers, R. W. & Yamamoto, R. T. Feeding on non-host plants by partially maxillectomized tobacco hornworms (*Manduca sexta*: Lepidoptera: Sphingidae). *Florida Entomol.* **75**, 89–93 (1992).
16. de Boer, G. Plasticity in food preference and diet-induced differential weighting of chemosensory information in larval *Manduca sexta*. *J. Insect Physiol.* **12**, 17–24 (1993).
17. Schoonhoven, L. M. & Dethier, V. G. Sensory aspects of host plant discrimination by lepidopterous larvae. *Arch. Neder. Zool.* **16**, 497–530 (1966).
18. Schoonhoven, L. M. Sensitivity changes in some insect chemoreceptors and their effect on food selection behavior. *Proc. K. Ned. Akad. Wet. C* **72**, 491–498 (1969).
19. Städler, E. & Hanson, F. E. in *Host Plant in Relation to Insect Behavior and Reproduction* (ed. Jermy, T.) 267–273 (Plenum, New York, 1976).
20. Abisgold, J. D. & Simpson, S. J. The effect of dietary protein levels and haemolymph composition on the sensitivity of maxillary palp chemoreceptors of locusts. *J. Exp. Biol.* **135**, 215–229 (1988).
21. Glendinning, J. L., Ensslen, S., Eisenberg, E. & Weiskopf, P. Diet-induced plasticity in the taste system of an insect: localization to a single transduction pathway in an identified taste cell. *J. Exp. Biol.* **202**, 2091–2102 (1999).
22. Rothschild, M., Aplin, R. & Marsh, N. Toxicity induced in the tobacco horn-worm (*Manduca sexta* L.) (Sphingidae, Lepidoptera). *Nature* **280**, 487–488 (1979).
23. Gothilf, S. E. & Hanson, F. E. Technique for electrophysiologically recording from sensory organs of intact caterpillars. *Entomol. Exp. Appl.* **72**, 305–310 (1994).

Acknowledgements

We thank J. Ewer, K. Whitlock, D. Bodnar, C. Gilbert, P. Rivlin and J. Bestman for critical reviews of the manuscript; M. Haribal for preliminary NMR and MS; B. Johnson for equipment to analyse electrophysiological recordings; and J. Almadovar and M. Chu for help in the laboratory. This work was supported by grants from the NSF and the Binational Science Foundation to R.B., and the NSF to C.L.M.

Correspondence and requests for materials should be addressed to M.L.d.C. (e-mail: moliva@binghamton.edu).

Scalable architecture in mammalian brains

Damon A. Clark^{*†}, Partha P. Mitra[‡] & Samuel S.-H. Wang^{*}

^{*} Department of Molecular Biology and [†] Department of Physics, Princeton University, Princeton, New Jersey 08544, USA

[‡] Bell Laboratories, Lucent Technologies, 600 Mountain Avenue, Murray Hill, New Jersey 07974, USA

Comparison of mammalian brain parts has often focused on differences in absolute size^{1–3}, revealing only a general tendency for all parts to grow together². Attempts to find size-independent effects using body weight as a reference variable¹ obscure size relationships owing to independent variation of body size⁴ and give phylogenies of questionable significance⁵. Here we use the brain itself as a size reference to define the cerebrototype, a species-by-species measure of brain composition. With this measure, across many mammalian taxa the cerebellum occupies a constant fraction of the total brain volume (0.13 ± 0.02), arguing against the hypothesis that the cerebellum acts as a computational engine

principally serving the neocortex³. Mammalian taxa can be well separated by cerebrototype, thus allowing the use of quantitative neuroanatomical data to test evolutionary relationships. Primate cerebrotypes have progressively shifted and neocortical volume fractions have become successively larger in lemurs and lorises, New World monkeys, Old World monkeys, and hominoids, lending support to the idea that primate brain architecture has been driven by directed selection pressure⁴. At the same time, absolute brain size can vary over 100-fold within a taxon, while maintaining a relatively uniform cerebrototype. Brains therefore constitute a scalable architecture.

Components of the brain make many more connections to one another than to any external structure. Furthermore, body size is under separate developmental and environmental control, rendering it an approximate reference measure at best⁴. Comparison of brain components across species, therefore, would be facilitated by a size measure that is independent of body parameters. To address this need for an internal normalization, we calculated the volume fraction (F) for each component, defining the total brain volume within each species as 1. We called the set of all the volume fractions for a species the cerebrototype.

A database of insectivores and primates⁶ showed significant variation of the cerebrototype (Fig. 1a, c). In insectivores, F did not vary systematically with brain size for any principal developmental brain division (Fig. 1a), therefore providing a baseline trend for comparison. By contrast, primates showed a strong trend for telencephalic growth¹. F_{telen} (volume fraction of telencephalon) was $60 \pm 4\%$ (mean $\pm$ s.d.; $n = 28$ species) in insectivores and $61 \pm 1\%$ in Scandentia (tree shrews; $n = 3$) but increased to $74 \pm 5\%$ in primates ($n = 44$). This increase occurred largely at the expense of medulla, mesencephalon and diencephalon: $F_{\text{med}} + F_{\text{mesen}} + F_{\text{dien}}$, respectively, was $27 \pm 3\%$ in insectivores and $26 \pm 1\%$ in tree shrews, and decreased to $14 \pm 4\%$ in primates. This trade-off was emphasized by the ratio $F_{\text{telen}}/(F_{\text{med}} + F_{\text{mesen}} + F_{\text{dien}})$, which was 2.3 ± 0.4 in insectivores, 2.3 ± 0.2 in tree shrews, 6.3 ± 3.2 in primates and 20.8 in *Homo sapiens*.

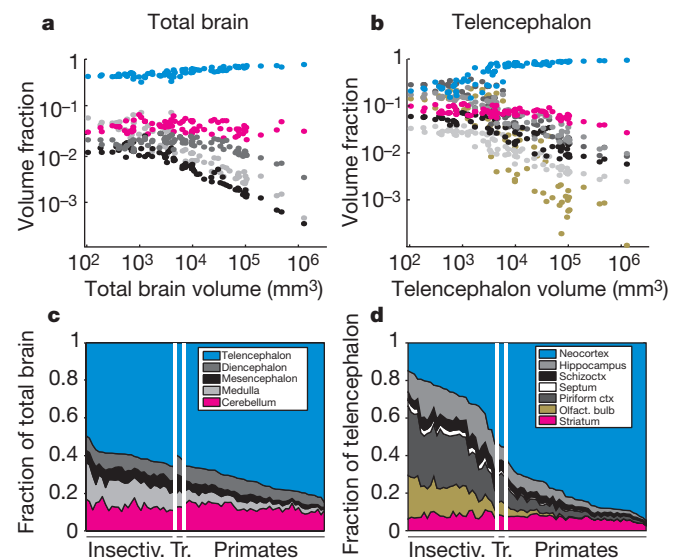


Figure 1 Analysis of volumetric data from mammalian brains. **a**, Volume fractions of principal brain divisions plotted against total brain volume. The colours are as indicated in **c**. **b**, Volume fractions of telencephalon subcomponents. The colours are as indicated in **d**. **c**, Area plot of principal brain-division volume fractions sorted by increasing telencephalic volume fraction within each taxon, demonstrating relative changes in non-cerebellar structures. Horizontal bars indicate the taxonomic groupings of insectivores (Insectiv.), tree shrews (Tr.) and primates. **d**, Area plot of telencephalic subcomponent volume fractions sorted by increasing neocortical volume fraction of the telencephalon.

Studies using body-weight-referenced scaling or comparison of absolute volumes³ suggested that cerebellum and neocortex vary together, leading to the proposition that the cerebellum is a computational engine that serves the needs of the neocortex. By contrast, we found that F_{cbl} (volume fraction of cerebellum) remained nearly constant across groups in the ref. 6 data set ($13.2 \pm 2.2\%$ in insectivores, $12.7 \pm 0.2\%$ in tree shrews and $12.4 \pm 2.1\%$ in primates; Fig. 1c, red points). To explore the generality of this finding, we examined F_{cbl} across 19 mammalian taxa (Fig. 2). F_{cbl} values across taxa remained constant ($13.5 \pm 2.4\%$) even as F_{neo} (volume fraction of neocortex) expanded from $16 \pm 6\%$ in Soricomorpha to $74 \pm 5\%$ in Hominoidea (apes and humans). Across this range of species, we found no correlation between F_{neo} and F_{cbl} ($r^2 = 0.013$, $n = 19$). The constancy of cerebellum in mammals suggests that its functional role must be evaluated not in terms of neocorticalization, but rather with either coordinated whole-brain function or some aspect of body plan that correlates strongly with total brain size.

F_{cbl} may be constrained by the high rate of energy consumption by brain tissue⁷, which would exert selection pressure for structures not to exceed a minimum essential size. Conversely, any taxon-specific increases in F_{cbl} might reflect necessary adaptations related to the addition of new functions. The cerebellum has been suggested to transform sensory information as a means of guiding motor activity. We found significantly higher values of F_{cbl} in mammals with sophisticated echolocating abilities⁸: Cetacea^{9–12} (dolphins and whales, $19 \pm 3\%$, 7 species; different from F_{cbl} in other taxa by two-tailed significance test, $P < 0.02$) and Microchiroptera¹³ (microbats, $22 \pm 5\%$, 225 species; different from other taxa, $P < 0.01$). In contrast to Microchiroptera, the Megachiroptera (megabats), which lack the ability to echolocate, did not have unusually large cerebella ($F_{\text{cbl}} = 14 \pm 1\%$, 47 species; $P = 0.97$), suggesting that adaptations to bat body plan alone were not responsible for the cerebellar enlargement. The observed size differences may be region specific: Cetacea and Microchiroptera are hypertrophied in specific

homologous parafloccular and medial regions, which in bats are responsive to acoustic target position and velocity (reviewed in ref. 14). Similarly, among teleost fishes, the largest cerebella are found in mormyrids, which electrolocate to detect object size and distance¹⁴. The common factor of echolocation and electrolocation is the interpretation of subtle timing differences in echoes returned from transmitted pulses. Thus, regardless of its overall functional role, the cerebellum seems in these species to provide additional processing power for sophisticated sensory adaptations. Our results also suggest a more general speculation: that the cerebellum provides guidance to the entire brain on the basis of fine features of sensory input.

Our observations in the whole brain led us to search for size trade-offs and constancy within the telencephalon. Normalizing to a total telencephalic volume of 1, telencephalic volume fractions T also showed high variability among orders (Fig. 1b, d). T_{neo} increased from $28 \pm 10\%$ ($n = 28$) in insectivores to $55 \pm 1\%$ ($n = 3$) in tree shrews, $81 \pm 8\%$ ($n = 44$) in primates and 95% in *Homo* (Fig. 1b, blue points). This expansion in neocortex occurred at the expense of hippocampus, septum, schizocortex, piriform cortex and olfactory bulb: the sum of these fractions decreased from $64 \pm 10\%$ in insectivores to $37 \pm 1\%$ in tree shrews, $12 \pm 7\%$ in primates and 3% in *Homo*.

In the face of this neocortical expansion, the telencephalic structure that showed the least change was the striatum (Fig. 1d, red points). The striatum maintained a volume fraction of $T_{\text{stri}} = 8 \pm 2\%$ in insectivores, $8 \pm 1\%$ in tree shrews, $6 \pm 1\%$ in primates and 3% in *Homo*. The striatum forms extensive interconnections with the hippocampus, the amygdala and nearly all of the cerebral cortex¹³. These connections form part of an extensive looped corticothalamic network that, when damaged, leads to extrapyramidal motor disorders such as Parkinson's and Huntington's disease. The relative preservation of T_{stri} is consistent with a function in motor activity that depends on the amount of telencephalic input and requires some minimum necessary amount of striatal processing.

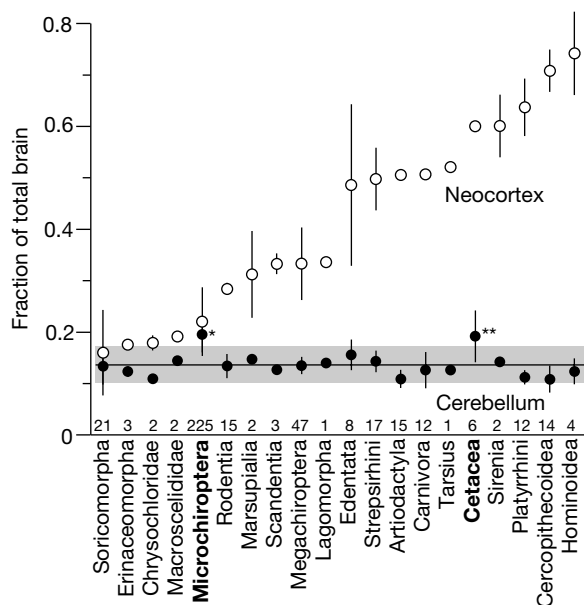


Figure 2 Constancy of cerebellar volume fraction across mammalian taxa. Cerebellar volume fractions are plotted at order level with the exception of the primates, which were divided into Strepsirhini (lemurs and lorises), Platyrrhini (New World monkeys), Cercopithecoidea (Old World monkeys) and Hominoidea (apes and human). The cerebellar volume fractions are indicated by solid symbols. Data are sorted by neocortical volume fraction (open symbols). The error bars indicate s.d.; the shaded bar indicates mean $\pm$ 1 s.d. for the pooled data. Statistically significant differences between an individual taxon and the entire group are identified in bold (Student's *t* test: asterisk, $P < 0.02$; double asterisk, $P < 0.01$).

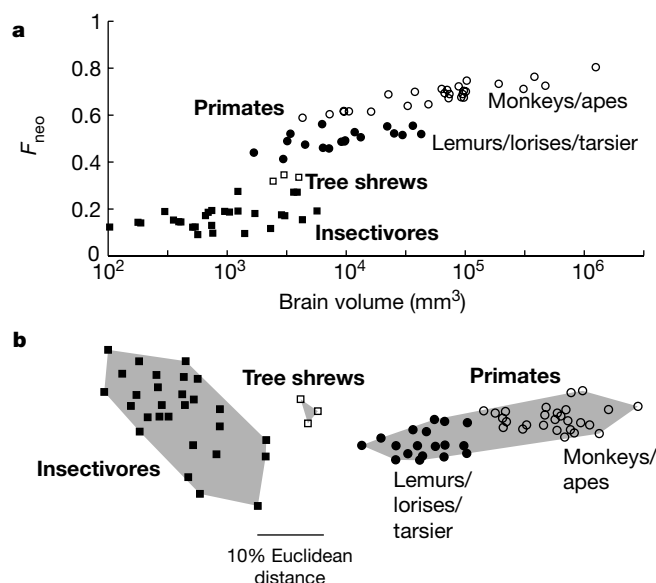


Figure 3 Clustering of cerebrotypes by taxon. **a**, Neocortical fraction of total brain volume plotted against total brain volume for insectivores (filled squares), tree shrews (open squares), lemurs, lorises and tarsier (filled circles), and monkeys and apes (open circles). **b**, Variation in the 11-component cerebrotypes for the species plotted in **a** displayed in a plane using multidimensional scaling. Compared with true Euclidean distance between cerebrotypes, the root-mean-squared (r.m.s.) fractional error in displayed interspecies distance is 8.4%.

The cerebrotypes we observe suggests that each taxon might be associated with its own distinct brain architecture. This can be seen at a coarse level by plotting a single quantity, F_{neo} , against brain volume¹. Although absolute brain volume varied several hundredfold within each order, there was little variation of F_{neo} (Fig. 3a), indicating that this fraction takes on specific values for major taxonomic divisions³.

To facilitate the identification of finer taxonomic distinctions, we used all total-brain and telencephalic volume fractions to generate a relational map of mammalian brain architectures. We did this using multidimensional scaling, an algorithm that shows the variation of all 11 volume fractions in a plane with maximum fidelity to the true distances between cerebrotypes. This method allows species to be compared without making *a priori* assumptions about allometric scaling relationships between components. Figure 3b shows clear separation between insectivores and primates. Scandentia, once categorized as primates⁶, were nearer to insectivores than to primates in cerebrotypes, but there was no overlap between either group. This assessment agrees with recent morphological and molecular evidence, and supports the identification of Scandentia as a separate mammalian lineage.

The eight 'insectivore'¹⁵ orders represented in this data set also appeared as distinct groups of points (Fig. 4a), with the exception of the overlapping cerebrotypes distributions of Soricidae (shrews) and Tenrecidae (tenrecs). This overlap arose entirely from the cerebrotypes of three tenrecs that live and/or hunt in the water (Fig. 4a). This difference from other tenrecs suggests that these three species

have brain specializations that reflect an aquatic lifestyle, such as reductions in the size of the olfactory bulbs and piriform cortex (see Supplementary Information). Thus, cerebrotypes analysis provides a means of taxonomically grouping mammals by their brain structure. The existence of these characteristic within-taxon cerebrotypes across large variations in absolute brain size suggests that brain components are scalable; that is, a functionally optimized brain has the same proportions independent of absolute volume. Furthermore, the fact that 11-component cerebrotypes can be mapped into a two-dimensional plane but still accurately represent distance information suggests that the range of possible brain architectures is strongly constrained.

Cerebrotypes-based measures might also be useful in finding directional changes in brain architecture. According to the social intelligence hypothesis of primate brain evolution⁴, a larger neocortex may confer selection advantages owing to increased cognitive capacity for use in social dynamics. This would lead to progressive, relative enlargement of the neocortex over time. Fossil and molecular evidence and morphological evidence from non-brain characters shows that successively more derived primate taxa have arisen over time: lemurs/orises, followed by tarsiers, New World monkeys, Old World monkeys, and then hominoids¹⁶. The arrangement of the corresponding cerebrotypes (Fig. 4b) is collinear with this order of primate taxa, with the exception of some overlap between New World and Old World monkeys. This overlap arises from the four New World monkeys in this database—*Ateles geoffroyi*, *Lagothrix lagotricha*, *Cebus* sp. and *Saimiri sciureus*—with larger group sizes and complex social structures resembling those of Old World monkeys¹⁷. The cerebrotypes of New World monkeys with large group sizes differ from other New World monkeys principally in F_{neo} (large group size $69.3 \pm 0.5\%$, 4 species; other $61.8 \pm 1.7\%$, 8 species; $P < 0.001$) but are similar to Old World monkeys ($70.4 \pm 2.4\%$, 11 species; $P = 0.1$; see Supplementary Information). Taken together, these orderings suggest that increased derivation was indeed accompanied by progressive changes in the cerebrotypes, perhaps relating to social intelligence.

Cerebrotypes shifts in primates are accompanied by significant expansion of the neocortical volume fraction¹ (Fig. 3a). To identify other, independent changes in architecture, we eliminated neocortex from the data set and renormalized the remaining ten regions. Multidimensional scaling still showed clear separation of primate taxa (Fig. 4c). We obtained similar results when we also removed the cerebellum from the data set (not shown). Thus, the rise of new primate taxa has been accompanied by substantial shifts in brain architecture, marked both by increases in relative neocortical size and by distinct changes in the distribution of brain volume among other regions.

Because cerebrotypes shifts are progressive, cerebrotypes differences should be largest between groups with the oldest last common ancestor. We emphasized changes in forebrain structures by considering the telencephalic cerebrotypes. At principal points of divergence for which dates have been estimated multiple times from DNA and fossil evidence¹⁶ (Fig. 5a), the Euclidean cerebrotypes distance was calculated between groups of species on either side of each node (see Methods) (Fig. 5b). The most recent divergence points were consistently associated with shorter distances between cerebrotypes (Spearman's rank correlation coefficient $r_s = 1.00$, $P < 0.01$). By contrast, a previous metric of body-size-scaled indices³ has suggested homology between the brains of humans and spider monkeys, a result irreconcilable with known phylogeny. Therefore, among quantitative brain parameters examined to date, only the cerebrotypes provides a measure of architecture that correlates with date of divergence of advanced primates.

To test if changes in brain architecture tracked primate evolution within each major group, the method of independent contrasts (see Methods)¹⁸ was used to make comparisons among groups of species (Fig. 5c). Overall, longer cerebrotypes distances were associated

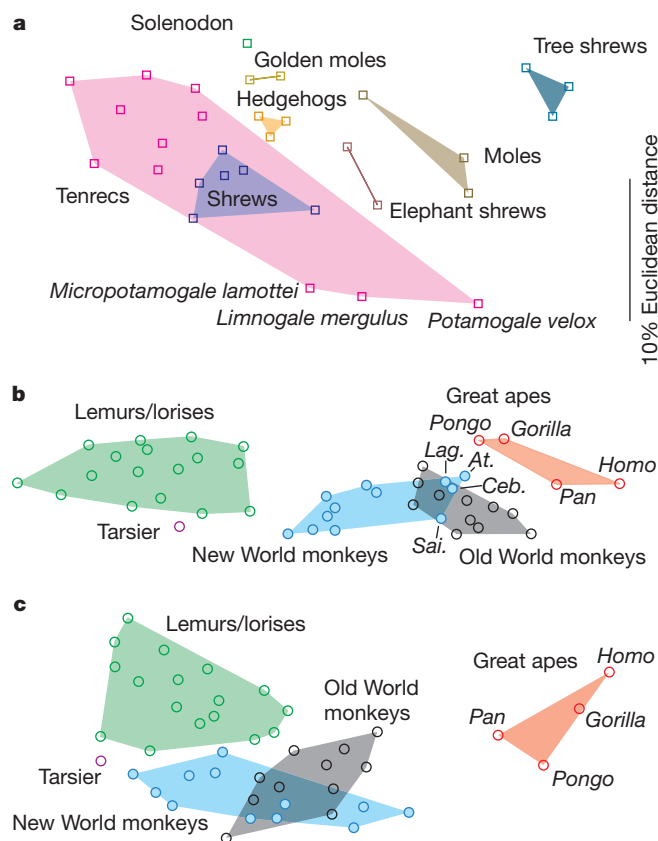


Figure 4 Clustering of cerebrotypes within taxa. **a**, Sorting of insectivore and tree shrew taxa using multidimensional scaling. Each shaded area indicates the minimum convex polygon for that taxon. The r.m.s. fractional error is 16.9%. **b**, Multidimensional scaling of primates. The r.m.s. fractional error is 6.3%. The labels for New World monkeys with complex social groupings indicate *A. geoffroyi* (At.), *L. lagotricha* (Lag.), *Cebus* sp. (Ceb.) and *S. sciureus* (Sai.). **c**, Multidimensional scaling of primates on cerebrotypes data in which neocortex was excluded and the remaining ten volume fractions renormalized. The r.m.s. fractional error is 14.2%. The scale bar in **a** also refers to **b** and **c**.

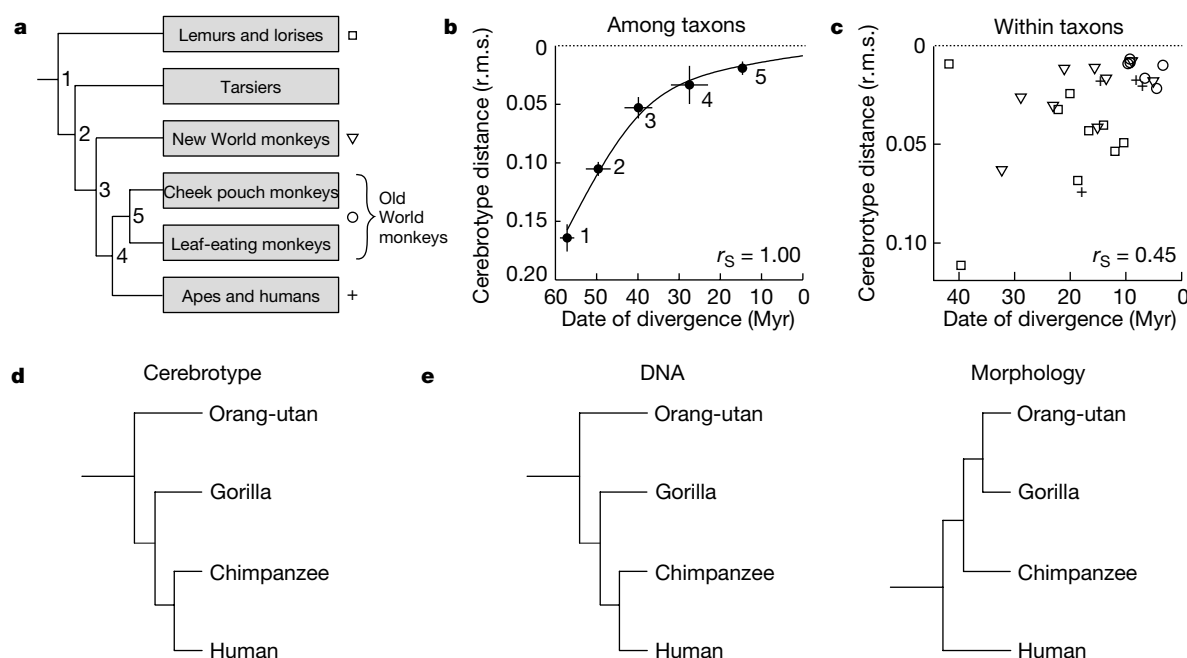


Figure 5 Evolutionary trends in brain architecture. **a**, Known phylogenetic relationships among primate taxa¹⁶ showing major divergence points (1–5). **b**, Telencephalic cerebrotypic distance between groups on either side of divergence points that are well-dated with fossil and DNA evidence. Vertical error bars indicate standard error of the mean; horizontal error bars are from ref. 16. The solid curve is a spline fit. **c**, Telencephalic

cerebrotypic distance within each taxon. Different symbols represent independent contrasts¹⁸ made within different taxa, as indicated in **a**. **d**, Relationships among hominoids reconstructed using telencephalic structures. **e**, Trees derived from DNA sequence (centre) and bone and tooth structure¹⁹ (right). The branch lengths are arbitrary.

with older divergence dates (Spearman's rank correlation coefficient $r_s = 0.45$, $P < 0.01$, $n = 28$ independent comparisons). This strong correlation led us to use distances between individual species to reconstruct the taxonomy of four hominoids (Fig. 5d). The resulting tree was identical to that obtained from DNA sequence (Fig. 5e, left), (((*Homo*, *Pan*), *Gorilla*), *Pongo*), but not to a tree obtained from bone and tooth morphology (Fig. 5e, right), (*Homo*, (*Pan*, (*Pongo*, *Gorilla*))). This lends support to the suggestion that soft tissues represent a more accurate route than hard tissues to phylogenetic reconstruction¹⁹.

Our results suggest a model for primate brain evolution in which adaptive radiation within a taxon has generated changes in the absolute brain (and also body) size while preserving an approximately fixed cerebrotypic. Shifts in cerebrotypic occurred at infrequent (about 10 Myr) intervals coinciding with the rise of new taxa. These shifts were marked primarily by relative expansion of the telencephalon, especially the neocortex. This expansion has reached an extreme in *Homo*, which possesses the largest F_{neo} value known (0.80).

Our observations suggest a reconciliation of developmentalist² and adaptationist³ models of how brain architecture is determined. Within each taxon, brain regions are scalable, tending to maintain fixed proportionality of size to one another independent of absolute total brain volume. This suggests that, within a taxon, the development of multiple brain regions is governed by a common set of mechanisms². However, because the cerebrotypic varies from taxon to taxon, these developmental mechanisms must also be variable. This variation arises from genetic modification that leads to the appearance of new cerebrotypes. For instance, in primates, the enlargement of the neocortex may be caused by the action of additional rounds of cortical neurogenesis^{20,21}. In addition to these major shifts, variations in developmental rates in other brain regions can lead to limited adaptive radiation of the cerebrotypic within each taxon. These variations may be driven by ecological and/or social requirements³. Finally, the microscopic

composition of neural tissue (neurons, glia, axons and dendrites) varies across mammals²². An ultimate explanation of these changes in brain structure will, therefore, require a deeper understanding of both the causes for variation at a cellular level and the selective advantages provided by a changing brain architecture. □

Methods

For cerebrotypic analysis, data were taken from ref. 6, consisting of 28 insectivores¹⁵, 3 Scandentia (tree shrews), 18 strepsirrhine (lemur and loris) primates and 27 haplorhine (tarsier, New World and Old World monkey, and hominoid) primates. Definitions of primate species conformed to ref. 16 and definitions of brain regions conformed to ref. 6. Brain volumes ranged over four orders of magnitude. The volume fraction was defined for the five principal brain components and for the seven telencephalic components. Whole-brain volume fractions (F) were calculated for medulla oblongata, cerebellum, mesencephalon, diencephalon and telencephalon by dividing by total brain volume. Volume fractions were calculated for olfactory bulb, piriform cortex, septum, striatum, schizocortex, hippocampus and neocortex by dividing by either the whole-brain volume (fractions F) or by the telencephalic volume (fractions T). We excluded the accessory olfactory bulb, which is not present in many primates, from the analysis. We also excluded one haplorhine (*Callicebus moloch*) from the analysis because the components failed to sum to the whole in the original data. Additional data used to compare neocortical and cerebellar volume fractions across taxa (Fig. 2) were obtained from the literature for Artiodactyla^{23,24}, Carnivora²⁴, Cetacea^{9–12}, Edentata²⁵, Lagomorpha²⁴, Marsupialia²⁴, Microchiroptera and Megachiroptera¹³, Rodentia^{26,27}, Sirenia^{28,29}, and orang-utan³⁰. All values are given as mean $\pm$ s.d. unless otherwise noted.

Independent contrasts

We arranged species in a tree according to a published primate phylogeny¹⁶. At each node, two mean cerebrotypes were calculated, one for each branch emanating from the node. The Euclidean distance between the two means was used to make a comparison for that node, the method of independent contrasts¹⁸.

Phylogenetic reconstruction

The telencephalic cerebrotypic was defined as T , the vector of telencephalic volume fractions. For hominoids, a phylogenetic tree was constructed from cerebrotypic distances using the Fitch–Margoliash algorithm, which sequentially adds species to a test tree while minimizing the quantity $\sum \Sigma (o_{ij} - e_{ij})^2 / o_{ij}^2$, where o_{ij} is the Euclidean cerebrotypic distance between species i and j , and e_{ij} is the total length of the path joining the two species in the test tree. Colobinae (leaf-eating monkeys) were used as the outgroup to root the tree. Optimization and display were done using the PHYLIP software package running on a PC-DOS platform.

Multidimensional scaling

Multidimensional scaling was done using MATLAB (The Mathworks, Natick, MA) to allow simultaneous display of variation in all 11 F components (principal brain divisions and telencephalic components) of the cerebrotypes. The quantity $\sum (o_{ij} - d_{ij})^2 / o_{ij}^2$ was minimized where o_{ij} is the cerebrototype distance and d_{ij} is the displayed distance in the plane. Minimizations were done using a bootstrapping procedure. The plot was initially seeded using three points chosen by a trial run to lie near the outskirts of the final plot. We added subsequent points one at a time in a random order. After each point was added, a round of error minimization was performed in which only the new point was allowed to vary, followed by a round of minimization in which all points were varied. Each minimization was performed at least ten times and the solution with the least error was accepted. In the 76-point minimization of all species (Fig. 3b), new points were added four at a time. This overall procedure resembles the Fitch–Margoliash algorithm for phylogeny reconstruction but yields a mapping in a plane rather than a connected tree.

Received 28 November 2000; accepted 12 March 2001.

- Jerison, H. J. *Brain Size and the Evolution of Mind* (American Museum of Natural History, New York, 1991).
- Finlay, B. L. & Darlington, R. B. Linked regularities in the development and evolution of mammalian brains. *Science* **268**, 1578–1584 (1995).
- Barton, R. A. & Harvey, P. H. Mosaic evolution of brain structure in mammals. *Nature* **405**, 1055–1058 (2000).
- Barton, R. A. & Dunbar, R. I. M. in *Machiavellian Intelligence II* (eds Whiten, A. W. & Byrne, R. W.) 240–263 (Cambridge Univ. Press, New York, 1992).
- Douglas, R. J. & Marcellus, D. The ascent of man: deductions based on a multivariate analysis of the brain. *Brain Behav. Evol.* **11**, 179–213 (1975).
- Stephan, H., Frahm, H. & Baron, G. New and revised data on volumes of brain structures in insectivores and primates. *Folia Primatol.* **35**, 1–29 (1981).
- Aiello, L. C. & Wheeler, P. The expensive-tissue hypothesis: the brain and digestive system in primate evolution. *Curr. Anthropol.* **36**, 199–221 (1995).
- Airapet'yants, E. S. & Konstantinov, A. I. *Echolocation in Animals* (Keter, Jerusalem, 1973).
- Jansen, J. On the whale brain with special reference to the weight of the fin whale (*Balaenoptera physalus*). *Norsk Hvalfangst-tidende* **9**, 480–486 (1952).
- Pilleri, G. Morphologie des gehirnes des "Southern Right Whale". *Acta Zool.* **46**, 245–272 (1964).
- Jansen, J. & Jansen, J. K. S. in *The Biology of Marine Mammals* (ed. Andersen, H. T.) 175–252 (Academic, New York, 1969).
- Ridgway, S. H. in *The Bottlenosed Dolphin, Tursiops spp.* (eds Leatherwood, J. S. & Reeves, R.) 69–97 (Academic, San Diego, 1989).
- Baron, G., Stephan, H. & Frahm, H. D. *Comparative Neurobiology in Chiroptera: Macromorphology, Brain Structures, Tables, and Atlases* (Birkhauser, Basel, 1996).
- Paulin, M. G. The role of the cerebellum in motor control and perception. *Brain Behav. Evol.* **41**, 39–50 (1993).
- Mouchaty, S. K., Gullberg, A., Janke, A. & Arnason, U. The phylogenetic position of the Talpidae within Eutheria based on analysis of complete mitochondrial proteins. *Mol. Biol. Evol.* **17**, 60–67 (2000).
- Purvis, A. A composite estimate of primate phylogeny. *Phil. Trans. R. Soc. Lond. B* **348**, 405–421 (1995).
- Kinzey, W. G. *New World Primates: Ecology, Evolution, and Behavior* (Aldine de Gruyter, New York, 1997).
- Harvey, P. H. & Pagel, M. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
- Gura, T. Bones, molecules...or both? *Nature* **406**, 230–233 (2000).
- Xuan, S. et al. Winged helix transcription factor BF-1 is essential for the development of the cerebral hemispheres. *Neuron* **14**, 1141–1152 (1995).
- Kornack, D. R. & Rakic, P. Changes in cell-cycle kinetics during the development and evolution of primate neocortex. *Proc. Natl Acad. Sci. USA* **95**, 1242–1246 (1998).
- Braitenberg, V. & Schüz, A. *Cortex: Statistics and Geometry of Neuronal Connectivity* (Springer, Berlin, 1998).
- Kruska, D. & Rohrs, M. Comparative-quantitative investigations on brains of feral pigs from the Galapagos Islands and of European domestic pigs. *Z. Anat. Entwicklungsgesch.* **144**, 61–73 (1974).
- Meyer, J. A quantitative comparison of the parts of the brains of two Australian marsupials and some eutherian mammals. *Brain Behav. Evol.* **18**, 60–71 (1981).
- Pirlot, P. & Kamiya, T. Quantitative brain organization in anteaters (Edentata-Tubilidentata). *J. Hirnforsch.* **24**, 677–689 (1983).
- Fox, J. H. & Wilczynski, W. Allometry of major CNS divisions: towards a reevaluation of somatic brain–body scaling. *Brain Behav. Evol.* **28**, 157–169 (1986).
- Frahm, H. D., Rehkämper, G. & Nevo, E. Brain structure volumes in the mole rat, *Spalax ehrenbergi* (Spalacidae, Rodentia) in comparison to the rat and subterranean insectivores. *J. Brain Res.* **38**, 209–222 (1997).
- Reep, R. L. & O'Shea, T. J. Regional brain morphometry and lissencephaly in the Sirenia. *Brain Behav. Evol.* **35**, 185–194 (1990).
- Pirlot, P. & Kamiya, T. Qualitative and quantitative brain morphology in the Sirenian Dugong dugong. *Erkl. Z. Zool. Syst. Evol.-forsch.* **23**, 147–155 (1985).
- Zilles, K. & Rehkämper, G. in *Orang-utan Biology* (ed. Schwartz, J. H.) 157–176 (Oxford Univ. Press, New York, 1988).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank J. M. Allman and M. J. Berry for discussions, R. Kasthuri for research assistance, and T. A. Barney for secretarial assistance. D.A.C. is in the Princeton University Program in Biophysics. S.S.-H.W. is supported by the Alfred P. Sloan Foundation.

Correspondence and requests for materials should be addressed to S.S.-H.W. (e-mail: samwang@molbio.princeton.edu).

An evolutionary scaling law for the primate visual system and its basis in cortical function

Charles F. Stevens

The Salk Institute and Howard Hughes Medical Institute,
10010 North Torrey Pines Road, La Jolla, California 92037, USA

A hallmark of mammalian brain evolution is the disproportionate increase in neocortical size as compared with subcortical structures¹. Because primary visual cortex (V1) is the most thoroughly understood cortical region, the visual system provides an excellent model in which to investigate the evolutionary expansion of neocortex. I have compared the numbers of neurons in the visual thalamus (lateral geniculate nucleus; LGN) and area V1 across primate species. Here I find that the number of V1 neurons increases as the 3/2 power of the number of LGN neurons. As a consequence of this scaling law, the human, for example, uses four times as many V1 neurons per LGN neuron (356) to process visual information as does a tarsier (87). I argue that the 3/2 power relationship is a natural consequence of the organization of V1, together with the requirement that spatial resolution in V1 should parallel the maximum resolution provided by the LGN. The additional observation that thalamus/neocortex follows the same evolutionary scaling law as LGN/V1 may suggest that neocortex generally conforms to the same organizational principle as V1.

Any study of evolutionary scaling relations—the allometric laws that relate the size of one structure to another—must deal with species that are homogeneous in their scaling properties. Various taxonomic orders or suborders can conform to scaling laws with the same power but different scale factors². The data presented here are for haplorhine primates, a suborder that appears to be homogeneous with respect to the brain areas that I consider.

Figure 1 plots the number of neurons in V1 as a function of number of LGN neurons for 23 haplorhines whose average brain volumes range from 3.4 cm³ (tarsier) to 1,252 cm³ (human). This figure is derived from data presented by various authors as indicated in the Methods. A nonlinear fit reveals that these data are well described by a power law with an exponent of 3/2 (1.54 ± 0.07). That is, across the haplorhines the number of V1 neurons N varies

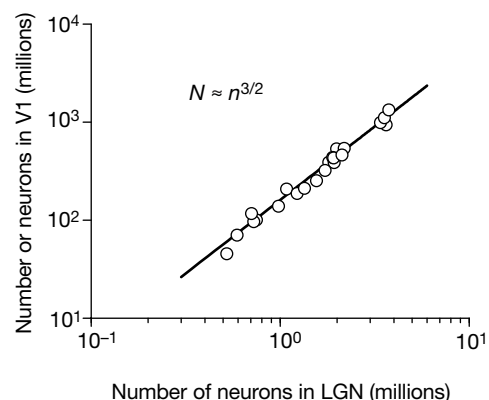


Figure 1 Number of neurons in V1 (N) as a function of number of LGN neurons (n) for 23 haplorhine primates. The smooth lines are power functions with exponents of 3/2 and 1.54 (the best fit to the data). For this graph, the tarsier has the smallest LGN/V1 ratio, and the human the largest.

with the number of LGN neurons n according to,

$$N = 161n^{3/2}$$

where numbers of neurons are measured in millions. For V1 and LGN, then, the number of neurons in neocortex increases with brain size more rapidly than the number of thalamic neurons, as has been observed¹; this increase follows an allometric relation with a power of $3/2$.

The $3/2$ scaling relation applies not only across species but, as shown in Fig. 2 (here represented as the volumes of V1 and LGN), within a single species (humans) for which the sizes of LGN and V1 co-vary between individuals by several fold³. The 23 haplorhines in Fig. 2 are the same as those appearing in Fig. 1, and the data were obtained from refs 3 and 4. The human data from ref. 3 were scaled so that the mean V1 and LGN volumes coincide with the human data point from ref. 4; this normalization was necessary to compensate for different measuring procedures used in the two studies.

How might this $3/2$ power law arise? The number of neurons in LGN is roughly the same as the number of retinal ganglion cells^{5,6}. This means that for larger animals with larger eyes and larger brains linear resolution of distances in the visual world increase as the $n^{1/2}$, where n is the number LGN neurons (close to the number of retinal ganglion cells and proportional to the number of visual pixels). To maintain the same spatial resolution in cortex, the number of cortical pixels must also increase in proportion to the number of LGN neurons. Cortical pixels are reasonably measured by the number of V1 hypercolumns⁷ or 'pinwheels'⁸, with each pinwheel being responsible for reporting on a small region of visual space; the orientation of a line in that part of the visual scene is specified by a 'location code', in other words, by which neurons in the pinwheel are responding. The pinwheel number should thus be proportional to the number of LGN neurons for visual scenes to be represented, at the cortical level, with the same resolution as provided by the eye. If the number of neurons per pixel were constant, then the number of V1 neurons would increase in proportion to the number of LGN neurons; but it does not.

How should the number of neurons per pixel (that is, per pinwheel) vary with retinal and LGN size, as measured by the number of LGN neurons? To answer this question, I must consider the uncertainty with which angles (θ) made by lines and edges (relative to the vertical) are known from a sample of the visual scene provided by the retina. A little thought shows that the angular resolution (which determines the uncertainty in angles) is proportional to the linear resolution. To see this, note that an angle is defined for a unit radius vector by the length of arc swept out by the vector, and that the resolution for this length is the linear resolution. Each pinwheel should, then, contain a number of neurons sufficient to maintain this angular resolution.

Pinwheels use a location code in which line (edge) orientation is represented by the identity of neurons in the pinwheel that are firing⁷. The angular resolution available in V1 is set by the number of neurons per pinwheel, and this must be proportional to $n^{1/2}$ for the angular resolution available from the retina to be preserved. A way to appreciate why this is so is to picture V1 as a three-dimensional space with dimensions representing x , y and θ . To have the same resolution in all directions, the number of neurons in each direction must be proportional to $n^{1/2}$. Altogether, the number of neurons in V1 should vary as the number of pixels (proportional to n) times the number of neurons per pixel (proportional to $n^{1/2}$): cortical neurons should increase as $n^{3/2}$, as is found experimentally.

This same argument can be made in a more abstract way. The two-dimensional visual representation in the retina and LGN (up/down—right/left) is mapped onto (and fills) a three-dimensional cortical space (up/down—right/left—line orientation). The volume of the three-dimensional cortical space must increase as the $3/2$ power of the area of two-dimensional retinal space as this 'input' space increases in size.

Of course V1 represents characteristics of the visual scene other than line orientation at each point in the visual world⁹. Characteristics such as ocular dominance and direction of movement appear, at least at the V1 level, to be represented by two populations of neurons (right eye/left eye, and opposite directions of edge movement) and these characteristics can be encoded by populations of neurons that are a constant fraction of the orientation-sensitive cells without losing information provided by the retinal ganglion cells. The population of neurons that represent colour can also increase in proportion to the number of orientation-sensitive cells without loss of information provided by the retina. Although the map of spatial frequencies may be continuous¹⁰, the argument used here would apply to this variable only if its resolution varied with brain size, which it probably does not. In general, the population of neurons that must increase like the $n^{1/2}$ are ones that make use of a location code to represent some continuous map characteristic (like line slope) whose precision must also be proportional to the resolution available in the map.

Does the argument given above apply to other areas of cortex and to other taxa? Unfortunately, similar data relating the number of thalamic neurons to number of cortical neurons (which that region of thalamus supplies) are not available for other cortical regions or for other taxa. Data are available, however, that relate thalamic volume to the volume of the entire neocortex^{4,11}, and these data are compared, for haplorhines, with the LGN/V1 volume data in Fig. 3. The slope on a double logarithmic plot of the overall thalamus/neocortex volumes is not significantly different from that of the LGN/V1 data, although the intercepts are different. This difference in intercepts presumably reflects the fact that primate V1 is 'two

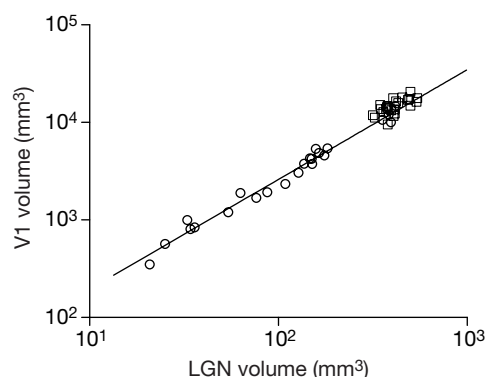


Figure 2 The volume of V1 as a function of LGN volume for 23 haplorhines (circles), together with data from 24 different human specimens (squares). The relation between cell number (Fig. 1) and the volumes of structures (this figure) is discussed in the Methods.

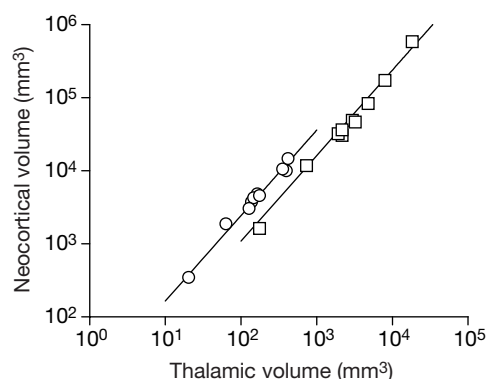


Figure 3 Volume of V1 (circles) and all of neocortex (squares) as a function of LGN and total thalamus volume. The data here are a subset, for which volume data are available, of the same 23 haplorhines as in Figs 1 and 2.

cortices in one', with twice the number of neurons underneath a square millimetre of surface than the remainder of neocortex¹². The density of neurons in the rest of neocortex scales similarly as in V1 (ref. 12), and all of the hominoid thalamic nuclei also have neuronal densities that vary in about the same way as in LGN^{13–15}; altogether, then, the relations between number of thalamic neurons and number of neocortical neurons is about the same as the 3/2 power scaling law described above for the primary visual cortex.

The conservation of these scaling relations raises the possibility that a similar basis for the scaling laws exists for all cortical areas. In this view, each cortical area would be provided with a map of some sort—perhaps one with very abstract quantities—and the job of the cortex would be to extract some characteristic of the map at each point that would be represented as a location code by the neurons in each map 'pixel'. Note that the information in the map need not be supplied by thalamus; this structure would only have to determine the number of pixels in the map. If n pixels are present in a cortical region, then the number of neurons per pixel needed to maintain the same resolution within a pixel as across pixels would vary as $n^{1/2}$. A 3/2 power relation would result. □

Methods

To obtain the relation between the number of LGN and V1 neurons, I must make use of three separately determined scaling relations. First, for haplorhines, the volume V of the grey matter in V1 is related to the LGN volume v by a power law^{4,16} (see Fig. 2),

$$V = Av^\alpha$$

where $A = 14.61 \pm 4.78$ and $\alpha = 1.125 \pm 0.057$; volumes are measured in cubic millimetres and refer to both hemispheres.

These volumes can be converted to numbers of neurons, if the neuronal densities in LGN and V1 are known. The numbers of neurons n in the LGN, as a function of LGN volume v , conform to a power law¹⁷ for 23 haplorhines and 17 strepsirhines,

$$n = Bv^\beta$$

with $\beta = 0.659 \pm 0.06$ for haplorhines and $\beta = 0.683 \pm 0.22$ for the strepsirhines, values that are not significantly different. The scale factor B , however, is different with a value of 0.071 ± 0.025 for haplorhines and 0.046 ± 0.029 for strepsirhines. This power law is based on data from ref. 17 combined with data from ref. 4.

The relation between V1 volume V and the number of V1 neurons N also follows the power law,

$$N = DV^\delta$$

where $D = 0.232$ and $\delta = 0.902$. This relation is obtained by using the observation¹² that 0.195 million neurons are found beneath a square millimetre of V1 surface in primates (the value is corrected for 18% shrinkage), and the weak power law dependence of V1 thickness t on cortical surface area S described^{12,18} by;

$$t = 0.825S^{0.108}$$

When V and v are converted to N and n with the equations above, a power law results (Fig. 1) with an exponent λ :

$$\lambda = \alpha\delta/\beta = 1.54 \pm 0.072$$

Received 7 December 2000; accepted 31 January 2001.

1. Finlay, B. L. & Darlington, R. B. Linked regularities in the development and evolution of mammalian brains. *Science* **268**, 1578–1584 (1995).
2. Barton, R. A. & Harvey, P. H. Mosaic evolution of brain structure in mammals. *Nature* **405**, 1055–1058 (2000).
3. Andrews, T. J., Halpern, S. D. & Purves, D. Correlated size variations in human visual cortex, lateral geniculate nucleus, and optic tract. *J. Neurosci.* **17**, 2859–2868 (1997).
4. Stephan, H., Frahm, H. D. & Baron, G. New and revised data on volumes and brain structures in insectivores and primates. *Folia primatol.* **35**, 1–29 (1981).
5. Schein, S. J. & de Monasterio, F. M. Mapping of retinal and geniculate neurons onto striate cortex of macaque. *J. Neurosci.* **7**, 996–1009 (1987).
6. Ahmad, A. & Spear, P. D. Effects of aging on the size, density, and number of rhesus monkey lateral geniculate neurons. *J. Comp. Neurol.* **334**, 631–643 (1993).
7. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (Lond.)* **195**, 215–243 (1968).
8. Bonhoeffer, T. & Grünwald, A. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* **353**, 429–431 (1991).
9. Swindale, N. V. How many maps are there in visual cortex? *Cereb. Cortex* **10**, 633–643 (2000).
10. Issa, N. P., Trepel, C. & Stryker, M. P. Spatial frequency maps in cat visual cortex. *J. Neurosci.* **20**, 8504–8514 (2000).
11. Frahm, H. D., Stephan, H. & Stephan, M. Comparison of brain structure volumes in Insectivora and Primates. I. Neocortex. *J. Hirnforsch.* **23**, 375–389 (1982).

12. Rockel, A. J., Hiorns, R. W. & Powell, T. P. The basic uniformity in structure of the neocortex. *Brain* **103**, 221–244 (1980).
13. Armstrong, E. A quantitative comparison of the hominoid thalamus. IV. Posterior association nuclei—the pulvinar and lateral posterior nucleus. *Am. J. Phys. Anthropol.* **55**, 369–383 (1981).
14. Armstrong, E. A quantitative comparison of the hominoid thalamus. II. Limbic nuclei anterior principalis and lateralis dorsalis. *Am. J. Phys. Anthropol.* **52**, 43–54 (1980).
15. Armstrong, E. Quantitative comparison of the hominoid thalamus. I. Specific sensory relay nuclei. *Am. J. Phys. Anthropol.* **51**, 365–382 (1979).
16. Frahm, H. D., Stephan, H. & Baron, G. Comparison of brain structure volumes in insectivora and primates. V. Area striata (AS). *J. Hirnforsch.* **25**, 537–557 (1984).
17. Shulz, H.-D. *Metrische Untersuchungen an den Schichten des Corpus Geniculatum Laterale tag- und Nachtaktiven Primaten*. Thesis, Johann Wolfgang Goethe-Universität Frankfurt (1967).
18. Fritschy, J. M. & Garey, L. J. Quantitative changes in morphological parameters in the developing visual cortex of the marmoset monkey. *Brain Res.* **394**, 173–888 (1986).

Acknowledgements

Much of the work described here was done at the Santa Fe Institute and the Aspen Center for Physics, and I thank those institutions for their support. I also thank T. Albright, S. Gandhi and I. Brivanlou for comments on an earlier draft.

Correspondence and requests for materials should be addressed to the author (e-mail: cfs@salk.edu).

Interocular rivalry revealed in the human cortical blind-spot representation

Frank Tong* & Stephen A. Engel†

* Department of Psychology, Princeton University, Princeton, New Jersey 08544, USA

† Department of Psychology, University of California Los Angeles, Los Angeles, California 90095, USA

To understand conscious vision, scientists must elucidate how the brain selects specific visual signals for awareness. When different monocular patterns are presented to the two eyes, they rival for conscious expression such that only one monocular image is perceived at a time^{1,2}. Controversy surrounds whether this binocular rivalry reflects neural competition among pattern representations or monocular channels^{3,4}. Here we show that rivalry arises from interocular competition, using functional magnetic resonance imaging of activity in a monocular region of primary visual cortex corresponding to the blind spot. This cortical region greatly prefers stimulation of the ipsilateral eye to that of the blind-spot eye. Subjects reported their dominant percept while viewing rivalrous orthogonal gratings in the visual location corresponding to the blind spot and its surround. As predicted by interocular rivalry, the monocular blind-spot representation was activated when the ipsilateral grating became perceptually dominant and suppressed when the blind-spot grating became dominant. These responses were as large as those observed during actual alternations between the gratings, indicating that rivalry may be fully resolved in monocular visual cortex. Our findings provide the first physiological evidence, to our knowledge, that interocular competition mediates binocular rivalry, and indicate that V1 may be important in the selection and expression of conscious visual information.

Despite extensive research, the neural basis of binocular rivalry has remained highly controversial. Specifically, it is debated whether discrepant monocular patterns rival because of interocular competition or pattern competition. Human psychophysical studies have provided evidence that rivalry results from interocular competition among monocular neurons in primary visual cortex (V1)³. However, single-unit recordings in awake, behaving monkeys have

yielded negligible evidence of rivalry-related activity in V1, inconsistent effects in visual areas V4 and MT, and strong effects in inferotemporal cortex^{4–6}. These neurophysiological findings instead indicate that rivalry may result from competition among incompatible pattern representations at higher levels of the visual pathway, well after inputs from the two eyes have converged in V1.

To resolve this issue, we used functional magnetic resonance imaging (fMRI) to monitor rivalry-related activity in a monocular region of human V1 corresponding to the blind spot. The blind spot is a part of the retina that has no photoreceptors; its size is around $4 \times 6^\circ$ and it is about 15° medial to the fovea (Fig. 1a). In human primary visual cortex, the blind spot is represented as a relatively large monocular region (around $10 \text{ mm} \times 5 \text{ mm}$; J. C. Horton, personal communication) that receives direct input solely from the ipsilateral eye and not from the (contralateral)

blind-spot eye. The monocular V1 blind-spot representation is large enough for functional imaging⁷.

Functional MRI has sufficient sensitivity and temporal resolution to detect rivalry-related responses in stimulus-selective extrastriate areas⁸. We predicted that if rivalry arises from interocular competition, then the ipsilateral-responsive neurons in the V1 blind-spot representation should show increased activity when subjects perceive a grating pattern presented to the ipsilateral eye and suppressed activity when subjects perceive a rivalrous grating presented to the blind-spot eye. Such excitation and suppression should occur even though both gratings are constantly present. Furthermore, if rivalry is fully resolved through interocular competition, then these neural responses during rivalry should be identical to those evoked by actual stimulus alternations between the ipsilateral grating and the blind-spot grating.

We performed three types of fMRI scan: V1 blind-spot localization scans, rivalry scans and stimulus alternation scans. Before MRI scanning, subjects mapped the visual field location of the right eye's blind spot by manipulating the location and size of a black flickering circular probe. During V1 blind-spot localization scans, subjects maintained fixation while viewing on/off sequences of a flickering checkerboard pattern using either their left ipsilateral eye or right blind-spot eye (Fig. 1a). The checkerboard was presented in the region of visual space corresponding to the blind spot and its immediate surround with a diameter of 8° , almost twice the size of the blind spot. Although the blind spot could not register the central portion of the checkerboard, the stimulus was perceptually filled in owing to stimulation of the blind spot's surround⁹.

The V1 blind-spot representation was reliably identified on the basis of the voxels in the left calcarine sulcus that showed a greater response to stimulation of the ipsilateral eye than of the blind-spot eye (see Methods). Figure 1 shows the fMRI response and anatomical locus of this region in one subject. The V1 blind-spot representation was highly monocular, responding vigorously to stimulation of the ipsilateral eye (Fig. 1b, left) and negligibly to stimulation of the blind-spot eye (Fig. 1b, right). In all subjects, this monocular region was located in the depth of the calcarine sulcus (Fig. 1c), a location that is always contained within primary visual cortex¹⁰.

During rivalry scans, a red vertical grating was presented to one eye and a green horizontal grating was presented to the other eye in the visual location corresponding to the blind spot and its surround

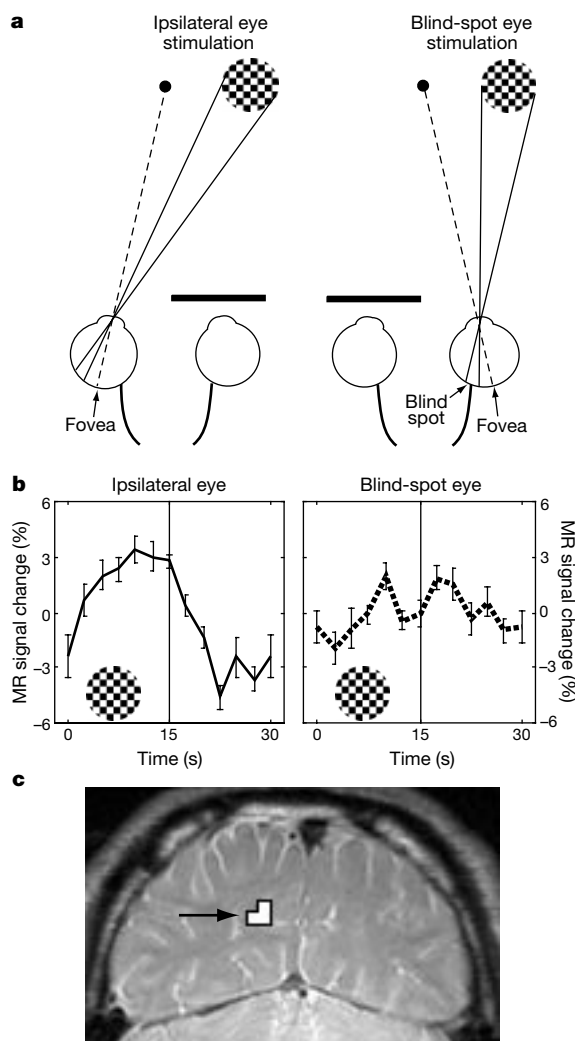


Figure 1 Localization of the V1 blind-spot representation. **a**, Viewing conditions. Subjects maintained fixation on a reference point while viewing a flickering checkerboard pattern (stimulus size 8° , check size 1° , temporal frequency 7.5 Hz, contrast 100%) with either the ipsilateral eye or blind-spot eye. The centre of the checkerboard fell on the blind spot (optic nerve head, size $\sim 4^\circ \times 6^\circ$) of the right eye but not the left eye. **b**, Average fMRI responses in the V1 blind-spot representation of one subject during stimulation of the ipsilateral or blind-spot eye. Data represent mean $\pm$ s.e. of eight stimulus periods (on 15 s, off 15 s) and are expressed in per cent signal change relative to the mean MR level throughout the scan. This region is activated by ipsilateral stimulation only. **c**, V1 representation of the right eye's blind spot in the left calcarine sulcus (three voxels highlighted in white, voxel size $3.1 \times 3.1 \times 4 \text{ mm}$, slice plane perpendicular to calcarine). Functional data are superimposed on high-resolution anatomical T2-weighted images.

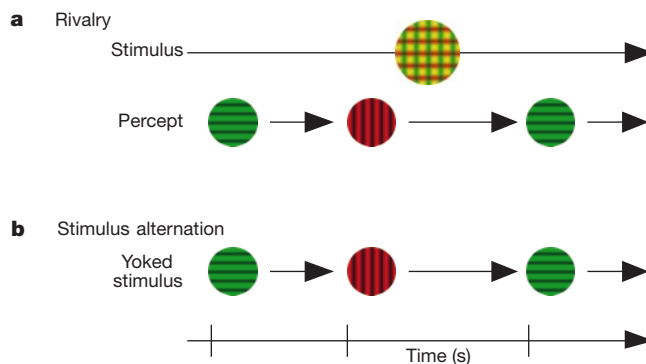


Figure 2 Binocular rivalry and stimulus alternation tasks. **a**, Rivalrous oscillating sine-wave gratings were presented to the blind-spot eye and ipsilateral eye (size 8°). When viewed through red and green filter glasses, only the green horizontal grating could be seen through one eye and only the red vertical grating through the other eye (spatial frequency 0.67 cycles per degree, speed 2 Hz, direction reversal every 500 ms, contrast 75%, mean luminance through matching filter 3.4 cd m^{-2}). Although both gratings were constantly present, subjects reported alternately perceiving either the red or green grating. **b**, On stimulus alternation scans, the physical stimulus alternated between the red grating and green grating using the sequence of reported alternations from a previous rivalry scan. Subjects reported when the stimulus changed to the red or green grating.

(Fig. 2a). Gratings moved back and forth within a stationary circular aperture to ensure continuous visual stimulation. Subjects used a button box to report whenever their dominant percept changed to the vertical grating or horizontal grating, or if a blend of the two gratings persisted.

On subsequent stimulus alternation scans, the physical stimulus alternated between non-rivalrous monocular presentations of either the red vertical grating or the green horizontal grating using the same sequence of alternations reported in a previous rivalry scan (Fig. 2b). To mimic the phenomenal alternations of rivalry, one monocular stimulus would gradually fade from maximum to zero contrast while the other monocular grating would gradually appear (zero to maximum contrast) over 250 ms. Subjects were instructed to report when the stimulus switched to the vertical grating, horizontal grating or a blend of the two stimuli.

Subjects reported normal rivalry alternations between the ipsilateral and blind-spot gratings with extensive periods of exclusive dominance and minimal perceptual blending (Table 1). The ability of the blind-spot surrounding grating to suppress the entire ipsilateral grating, including its central 'unpaired' region, is consistent with the finding that binocular rivalry can occur among nonoverlapping stimuli². In separate psychophysical studies, we have confirmed the rivalrous nature of these interactions encompassing the blind spot—increasing the contrast of either grating decreased the dominance duration of the opposing grating, as is found in foveal vision¹¹.

Three out of four subjects showed significantly longer dominance durations for the ipsilateral grating than for the blind-spot grating

(S1, S2, S4, $t > 2.0$, $P < 0.05$). These behavioural findings, though preliminary, are consistent with the hypothesis that rivalry dominance depends upon the ratio of monocular neurons activated by each eye³.

We calculated fMRI responses by averaging the fMRI time course surrounding all occurrences of a reported switch to the ipsilateral grating or blind-spot grating, time-locked to each reported switch. Figure 3 shows the average fMRI responses of each subject for rivalry versus stimulus alternation. The vertical bar at time zero indicates the time of the reported switch.

Although both gratings were constantly present during rivalry, the monocular blind-spot representation showed a sharp increase in fMRI activity soon after subjects reported that the ipsilateral grating had become perceptually dominant (Fig. 3a, green solid line). This rise in activity reflects the increased firing of the monocular neurons that receive input from the ipsilateral eye. Conversely, when the blind-spot grating became dominant, activity in this monocular region showed a sharp decrease (Fig. 3a, red dotted line). Thus, the signals from the ipsilateral eye to the V1 blind-spot representation were suppressed when the stimulus entering the other eye became perceptually dominant.

All four subjects showed the same qualitative pattern of awareness-related responses during rivalry. This tight correspondence between visual awareness and neural activity in monocular visual cortex confirms the predictions of interocular rivalry.

Functional MRI responses evoked by stimulus alternation (Fig. 3b) were remarkably similar to those observed during rivalry in their pattern, magnitude and timing. The initial peak or trough of the response always occurred from 0 to 2 s after the reported switch and the final peak or trough always occurred between 3 and 6 s after the switch. These peak-to-trough differences were highly reliable, yielding statistically significant linear or quadratic trends in fMRI responses within this time window (0 to 6 s) for all subjects, switch types and conditions ($F > 4.1$, $P < 0.05$).

To assess whether rivalry was fully resolved in the monocular V1 blind-spot representation, we compared the amplitude of fMRI responses (final minus initial peak-to-trough difference) for rivalry versus stimulus alternation. Figure 4 shows the normalized fMRI response amplitude of each subject, switch type and condition. Statistical analyses revealed that fMRI responses did not differ for

Table 1 Perceptual dominance durations reported during rivalry

Subject	Mean dominance duration (s)			Relative predominance (%)		
	Ipsilateral	Blind-spot	Blend	Ipsilateral	Blind-spot	Blend
S1	4.8	3.6	0.0	58	42	0
S2	2.9	2.3	0.7	55	41	4
S3	4.4	4.1	2.2	45	42	13
S4	6.1	2.8	1.4	65	29	6
Mean	4.6	3.2	1.1	56	38	6

Dominance durations for each grating followed a gamma-shaped distribution characteristic of binocular rivalry²; mean dominance durations are shown. Relative predominance is the percentage of total viewing time that the subject reported perceiving the ipsilateral grating only, the blind-spot grating only or a perceptual blend of both gratings.

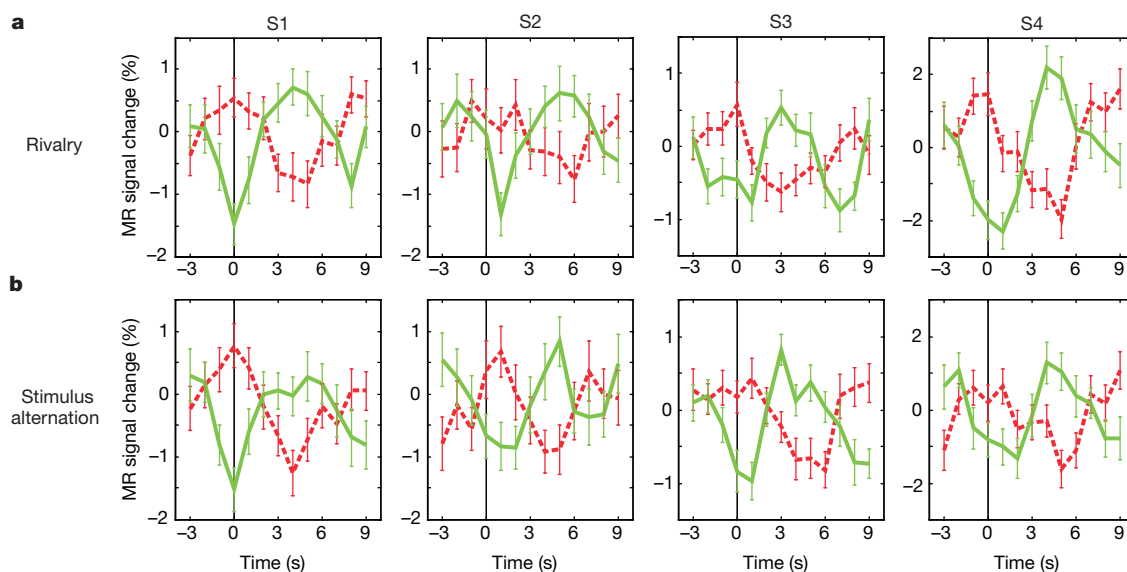


Figure 3 fMRI activity during rivalry and stimulus alternation. Average fMRI activity in the V1 blind-spot representation during perceptual switches to the ipsilateral grating (green solid line) or blind-spot grating (red dotted line) for rivalry (**a**) versus stimulus alternation (**b**). Data of all four subjects are plotted on individually scaled y-axes. Vertical lines at time zero indicate the time of the subject's response. **a**, During rivalry, fMRI activity increases

sharply soon after the ipsilateral grating becomes dominant in awareness and decreases when the blind-spot grating becomes dominant, consistent with the predictions of interocular competition. **b**, Very similar fMRI responses occur during stimulus alternations between the two monocular gratings. (fMRI responses typically peak 2–6 s after stimulus onset because of haemodynamic lag.)

rivalry versus stimulus alternation ($F < 1$). Positive responses for the ipsilateral grating were larger than negative responses for the blind-spot grating ($F = 102$, $P < 0.005$) but rivalry and stimulus alternation remained equivalent across both types of neural response.

The equivalence between fMRI responses for rivalry and stimulus alternation indicates that it is likely that rivalry has been entirely resolved among monocular neurons in the V1 blind-spot representation, such that neural activity entirely reflects the subject's perceptual state. Thus, functionally equivalent neural responses are observed when the subject's conscious state alternates between the ipsilateral grating and blind-spot grating during constant rivalrous stimulation and when the physical stimulus itself alternates between each grating shown alone.

Our results show that binocular rivalry is resolved in monocular visual cortex and provide physiological evidence to support interocular competition. These theories predict that left-eye versus right-eye inputs are alternately suppressed during rivalry because of lateral inhibition among monocular V1 neurons³ or feedback inhibition from V1 to monocular layers of the lateral geniculate nucleus¹². In contrast, theories of pattern competition propose that rivalry occurs among binocular pattern neurons at much higher levels of the visual pathway and not among monocular neurons⁴. Our findings therefore help to resolve the neural basis of binocular rivalry.

Previous studies have found evidence of rivalry-related neural activity but none has established the involvement of monocular neurons^{4–6,8,13–18}. In a single-unit study in monkeys, only 3 out of 33 V1 neurons showed significant responses corresponding to conscious perception, suggesting that rivalry takes place at higher levels of the visual pathway⁴. However, reanalysis revealed that across this V1 population, responses during rivalry equalled one-third of the magnitude of stimulus alternation responses¹³. These results are more consistent with an fMRI study showing reliable rivalry responses in human V1 that were about half the magnitude of stimulus alternation responses¹³. (Unfortunately, this study could not isolate monocular responses.) The authors suggested many factors that might account for the stronger rivalry effects in human V1, including interspecies differences, the indirect nature of fMRI in estimating neural activity, and the effects of eye movements on single-unit recordings.

In our view, the present finding of equally powerful rivalry and

stimulus alternation responses strongly suggests that binocular rivalry is resolved in monocular visual cortex. Although fMRI provides an indirect estimate of neural activity, any factors that might inflate response amplitudes during rivalry would also do so during stimulus alternation. In contrast, certain factors may have diluted rivalry responses in previous studies. These include sub-optimal viewing conditions that lead to frequent perceptual blends, and variability in the accuracy or timing of subjects' perceptual report. Isolating factors that weaken rivalry responses is an important direction for future research.

Although competition among binocular pattern neurons alone cannot account for our findings, it remains possible that feedback signals from binocular neurons to monocular neurons might yield the interocular suppression that we observe. However, such a theory fails to explain why pattern competition should lead to selection at the monocular level. Furthermore, it remains unclear how feedback projections from binocular neurons might target a specific monocular channel. Given these difficulties, interocular competition provides the most compelling explanation for rivalry in monocular visual cortex.

Our data also address a debate regarding whether common or separate neural mechanisms underlie binocular rivalry and related phenomena involving pattern rivalry. For example, two low-contrast patterns presented to one eye can weakly rival with each other^{1,19}. Moreover, one of two dichoptic patterns can maintain dominance even when the patterns are frequently swapped between eyes^{20,21}. Such rivalry probably involves high-level pattern competition. One proposal is that pattern competition may generally account for binocular rivalry^{20,22}. However, our results suggest that a separate mechanism of interocular competition entirely accounts for the binocular rivalry in our subjects.

Finally, our findings show that neurons can reflect conscious perception at a much earlier level of the visual pathway than previously thought^{4,23}. These results have one of two implications. One possibility is that certain aspects of conscious vision begin to emerge at the very earliest stage of cortical processing among monocular V1 neurons. Alternatively, our findings may suggest a new role for V1 as the 'gatekeeper' of consciousness, a primary cortical region that can select which visual signals gain access to awareness. In either case, our study firmly establishes the importance of primary visual cortex in binocular rivalry and conscious vision. □

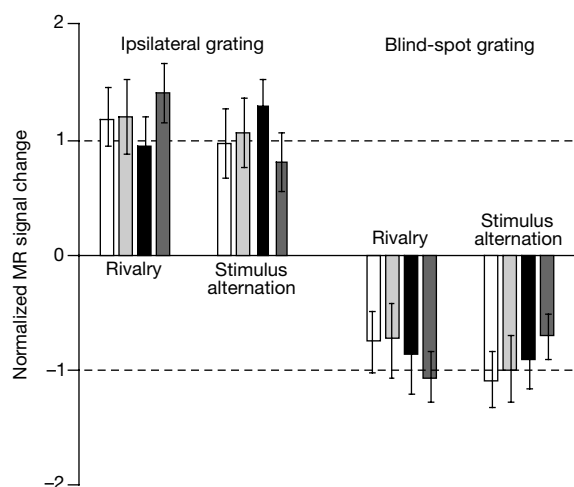


Figure 4 fMRI response amplitudes for rivalry versus stimulus alternation. Amplitude of fMRI responses (peak-to-trough difference, time window 0 to +6 s) of each subject ($n = 4$) for switches to the ipsilateral grating (left) and blind-spot grating (right) during rivalry versus stimulus alternation. Responses are normalized relative to each subject's mean response magnitude across the four conditions. fMRI responses for rivalry versus stimulus alternation did not reliably differ ($F < 1$).

Methods

Subjects

Four healthy right-handed volunteers (three women), aged 21–28, participated. All had normal or corrected-to-normal visual acuity and normal stereo-depth perception. Two subjects were right-eye dominant and two were left-eye dominant. Before the fMRI experiment, subjects received training on how to localize their blind spot in visual space and to report their online perception during binocular rivalry and stimulus alternation.

fMRI acquisition

Subjects were scanned at the UCLA Division of Brain Mapping on a 3T General Electric scanner using a head coil. We collected functional images using 5–6 slices oriented perpendicular to the calcarine sulcus with the first slice beginning about 10 mm anterior to the occipital pole (slice thickness 4 mm, inter-slice distance 1 mm, in-plane resolution 3.125×3.125 mm). We used standard T2*-weighted echoplanar imaging to localize the V1 blind-spot representation ($TR = 2.5$ s, $TE = 45$ ms, flip angle 80°) and for rivalry and stimulus alternation scans ($TR = 1.0$ s, $TE = 45$ ms, flip angle 45°). Functional images were superimposed on high-resolution T2-weighted anatomical images (in-plane resolution 0.78×0.78 mm). A bite bar minimized head motion.

V1 blind-spot localization

On separate fMRI scans, subjects viewed a counterphasing black/white checkerboard pattern with either their left ipsilateral eye or right blind-spot eye as shown in Fig. 1a. (In the actual experiment, a real fixation point was placed around 15° to the left of the subject's midline and the stimulus appeared roughly on the midline at a distance that closely corresponded to the subject's horopter.) Each scan consisted of an initial 30-s rest period followed by eight cycles of stimulation (15 s) and rest (15 s). We calculated activation maps and difference maps using multiple regression²⁴. A linear model consisting of sinusoidal response functions was fit to the fMRI activity observed during the eight cycles of visual

stimulation for each eye. We identified the V1 blind-spot representation in individual subjects as those voxels in the left calcarine sulcus that showed both a significant response to ipsilateral stimulation and a significantly greater response to ipsilateral than blind-spot stimulation using a minimum statistical threshold of $t = 2.0$, $P < 0.05$. The blind-spot representation ranged in size from 3–5 voxels (voxel size $3.125 \times 3.125 \times 4$ mm) across subjects, consistent with size estimates based on post-mortem neuroanatomical studies (J. C. Horton, personal communication).

Binocular rivalry and stimulus alternation scans

During these scans, two subjects viewed the red vertical grating and green horizontal grating with their left eye and right eye, respectively, whereas two subjects received the reverse eye assignment. Subjects performed 7–10 scans of rivalry and an equal number for stimulus alternation. Each scan lasted for 90 s. We discarded the first 10 s of fMRI activity to remove transient responses to the onset of the stimulus. We converted fMRI activity from the V1 blind-spot representation to per cent signal change from the mean level during the scan, and potential MR spikes and artefacts were minimized by reducing any outliers to lie within 3 s.d. of the mean.

We conducted an event-related fMRI analysis for reported switches between the blind-spot and ipsilateral grating. Previously, we found that rivalry responses increase as a function of percept duration and that very brief percepts led to unreliable fMRI responses⁸. Here, a switch was considered valid only if the percept immediately preceding and following the reported switch lasted longer than 2 s. An intervening blend response was allowed if it occurred within 1 s before the reported switch, in which case the blend duration was incorporated into the pre-switch period. fMRI responses of each subject were calculated by separately averaging the fMRI time course surrounding all occurrences of a reported switch to the ipsilateral grating or blind-spot grating for rivalry versus stimulus alternation, time-locked to each reported switch (rounded to the nearest second). Each average fMRI response function consisted of 39–55 observations.

Received 14 December 2000; accepted 12 February 2001.

- Helmholtz, H. v. *Helmholtz's Treatise on Physiological Optics* (The Optical Society of America, Rochester, New York, 1924).
- Levelt, W. J. M. *On Binocular Rivalry* (Royal VanGorcum, Assen, The Netherlands, 1965).
- Blake, R. A neural theory of binocular rivalry. *Psychol. Rev.* **96**, 145–167 (1989).
- Leopold, D. A. & Logothetis, N. K. Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature* **379**, 549–553 (1996).
- Logothetis, N. K. & Schall, J. D. Neural correlates of subjective visual perception. *Science* **245**, 761–763 (1989).
- Sheinberg, D. L. & Logothetis, N. K. The role of temporal cortical areas in perceptual organization. *Proc. Natl Acad. Sci. USA* **94**, 3408–3413 (1997).
- Tootell, R. B. H. et al. Functional analysis of primary visual cortex (V1) in humans. *Proc. Natl Acad. Sci. USA* **95**, 811–817 (1998).
- Tong, F., Nakayama, K., Vaughan, J. T. & Kanwisher, N. Binocular rivalry and visual awareness in human extrastriate cortex. *Neuron* **21**, 753–759 (1998).
- Walls, G. L. The filling in process. *Am. J. Optom. Arch. Am. Acad. Optom.* **31**, 329–341 (1954).
- Stensaas, S. S., Eddington, D. K. & Doherty, W. H. The topography and variability of the primary visual cortex in man. *J. Neurosurg.* **40**, 747–755 (1974).
- Fox, R. & Rasche, F. Binocular rivalry and reciprocal inhibition. *Percept. Psychophys.* **5**, 215–217 (1969).
- Lehky, S. R. An unstable multivibrator model of binocular rivalry. *Perception* **17**, 215–228 (1988).
- Polonsky, A., Blake, R., Braun, J. & Heeger, D. J. Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature Neurosci.* **3**, 1153–1159 (2000).
- Lansing, R. W. Electroencephalographic correlates of binocular rivalry in man. *Science* **146**, 1325–1327 (1964).
- Cobb, W. A., Morton, H. B. & Ettlinger, G. Cerebral potential evoked by pattern reversal and their suppression in visual rivalry. *Nature* **216**, 1123–1125 (1967).
- Brown, R. J. & Norcia, A. M. A method for investigating binocular rivalry in real-time with the steady-state VEP. *Vision Res.* **37**, 2401–2408 (1997).
- Tononi, G., Srinivasan, R., Russell, D. P. & Edelman, G. M. Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. *Proc. Natl Acad. Sci. USA* **95**, 3198–3203 (1998).
- Lumer, E. D., Friston, K. J. & Rees, G. Neural correlates of perceptual rivalry in the human brain. *Science* **280**, 1930–1934 (1998).
- Wade, N. J. Monocular and binocular rivalry between contours. *Perception* **4**, 85–95 (1975).
- Logothetis, N. K., Leopold, D. A. & Sheinberg, D. L. What is rivaling during binocular rivalry? *Nature* **380**, 621–624 (1996).
- Lee, S. H. & Blake, R. Rival ideas about binocular rivalry. *Vision Res.* **39**, 1447–1454 (1999).
- Andrews, T. J. & Purves, D. Similarities in normal and binocularly rivalrous viewing. *Proc. Natl Acad. Sci. USA* **94**, 9905–9908 (1997).
- Crick, F. & Koch, C. Are we aware of neural activity in primary visual cortex? *Nature* **375**, 121–123 (1995).
- Friston, K. J. et al. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1994).

Acknowledgements

We thank A. Seiffert and K. Nakayama for comments on this manuscript; D. Tran for research assistance; and J. Mazziotta, M. Cohen, the UCLA Brain Mapping Medical Organization, the Ahmanson Foundation, the Pierson-Lovelace Foundation, the Tamkin Foundation, and the Jennifer Jones-Simon Foundation for support. This research was funded by a McDonnell-Pew Grant in Cognitive Neuroscience and by the National Institutes of Health.

Correspondence and requests for materials should be addressed to F. T. (e-mail: ftong@princeton.edu).

Linkage disequilibrium in the human genome

David E. Reich*, Michele Cargill*, Stacey Bolk*, James Ireland*, Pardis C. Sabeti*, Daniel J. Richter*, Thomas Lavery*, Rose Kouyoumjian*, Shelli F. Farhadian*, Ryk Ward* & Eric S. Lander*§

* Whitehead Institute / MIT Center for Genome Research,

Nine Cambridge Center, Cambridge, Massachusetts 02142, USA

‡ Institute of Biological Anthropology, University of Oxford, Oxford OX2 6QS, UK

§ Department of Biology, MIT, Cambridge, Massachusetts 02139, USA

With the availability of a dense genome-wide map of single nucleotide polymorphisms (SNPs)¹, a central issue in human genetics is whether it is now possible to use linkage disequilibrium (LD) to map genes that cause disease. LD refers to correlations among neighbouring alleles, reflecting 'haplotypes' descended from single, ancestral chromosomes. The size of LD blocks has been the subject of considerable debate. Computer simulations² and empirical data³ have suggested that LD extends only a few kilobases (kb) around common SNPs, whereas other data have suggested that it can extend much further, in some cases greater than 100 kb^{4–6}. It has been difficult to obtain a systematic picture of LD because past studies have been based on only a few (1–3) loci and different populations. Here, we report a large-scale experiment using a uniform protocol to examine 19 randomly selected genomic regions. LD in a United States population of north-European descent typically extends 60 kb from common alleles, implying that LD mapping is likely to be practical in this population. By contrast, LD in a Nigerian population extends markedly less far. The results illuminate human history, suggesting that LD in northern Europeans is shaped by a marked demographic event about 27,000–53,000 years ago.

To characterize LD systematically around genes, each of the 19 regions that we studied was anchored at a 'core' SNP in the coding region of a gene. The core SNP was chosen from a database of more than 3,000 coding SNPs that had been identified by screening in a multi-ethnic panel (see Methods), subject to two requirements. First, 'finished' genomic sequence was available for 160 kb in at least one direction from the core SNP; second, the frequency of the minor (less common) allele was at least 35% in the multi-ethnic panel.

We focused on high-frequency SNPs for several reasons. First, they tend to be of high frequency in all populations⁷, facilitating cross-population comparisons. Second, LD around common alleles represents a 'worst case' scenario: LD around rare alleles is expected to extend further because such alleles are generally young⁸ and there has been less historical opportunity for recombination to break down ancestral haplotypes². Third, LD around common alleles can be measured with a modest sample size of 80–100 chromosomes to a precision within 10–20% of the asymptotic limit (see Methods). Last, LD around common alleles will probably be particularly relevant to the search for genes predisposing to common disease⁹.

To identify SNPs at various distances from the core SNP, we re-sequenced subregions of around 2 kb centred at distances 0, 5, 10, 20, 40, 80 and 160 kb in one direction from the core SNP using 44 unrelated individuals from Utah. Altogether, we screened 251,310 bp (see Methods) and found an average heterozygosity of $\pi = 0.00070$, consistent with past studies¹. A total of 272 'high frequency' polymorphisms were identified (Table 1).

We measured LD between two SNPs using the classical statistic D' (see Methods)¹⁰. D' has the same range of values regardless of the frequencies of the SNPs compared¹¹. Its sign (positive or negative)

† Present address: Celera Genomics, 45 West Gude Drive, Rockville, Maryland 20850, USA.

depends on the arbitrary choice of the alleles paired at the two loci. We chose the pair of SNPs that caused $D' > 0$ in Utah so that, in comparisons with other populations, the sign of D' indicates whether the same or opposite allelic association is present. In a large sample, $|D'|$ of 1 indicates complete LD; 0 corresponds to no LD. The degree of LD needed for effective mapping depends on the details of a particular study². A useful measure is the 'half-length' of

LD (the distance at which the average $|D'|$ drops below 0.5). Comparing the 19 randomly selected regions, LD has a half-length of about 60 kb (Fig. 1). Significant P -values for LD occur in greater than 50% of cases at distances of ≤ 80 kb. LD therefore extends much further than a previous prediction², and our data indicate that, in general, blocks of LD are large. Although the average extent is large, there is great variation in LD across the

Table 1 Distribution of regions and SNPs within regions

Gene identification*	Chromosome	Local recombination rate (cM Mb ⁻¹)†	Span of region in physical map (cR)‡	Number of high-frequency polymorphisms at distances (kb) from core SNPs§						
				0	5	10	20	40	80	160
BMP8l	1	1.4	60.88–61.04	1	1	1	2	2	1	3
ACVR2Bl	3	1.1	37.27–37.43	1	1	2	3	0	3	1
TGFB1	5	1.4	152.16–152.00	2	3	1	0	1	7	0
DDR1l	6	–	46.046–45.89	1	4	1	4	2	2	0
GTF2H4	6	–	46.059–46.22	1	3	3	6	2	0	0
COL11A2l	6	–	48.27–48.43	3	0	2	2	3	1	1
LAMB1l	7	2.3	106.58–106.42	1	0	5	3	3	2	4
WASL	7	0.5	122.99–122.83	0	1	0	4	2	2	1
SLC6A12	12	3.3	3.62–3.78	2	8	2	1	6	0	0
KCNA1	12	3.3	8.50–8.66	2	1	5	2	1	1	0
SLC2A3	12	2.1	16.33–16.49	1	2	1	1	5	0	2
ARHGDIBl	12	1.2	21.84–22	1	9	1	2	2	1	2
PCll	14	4.3	98.41–98.57	3	7	0	0	4	14	1
PRKCB1	16	1.0	32.50–32.66	0	2	3	0	4	3	0
NFI	17	1.0	38.10–38.26	0	0	2	2	5	1	0
SCYA2l	17	3.0	40.21–40.37	1	5	1	1	3	1	1
PAI2	18	2.7	64.17–64.01	0	0	2	1	5	2	10
IL17Rl	22	5.9	14.48–14.32	2	1	2	1	2	0	4
HCF2l	22	2.0	17.91–17.75	1	1	2	2	1	0	3

* Abbreviations from LocusLink (www.ncbi.nlm.nih.gov/LocusLink/list.cgi).
† For three regions the genetic and physical maps were inconsistent and no estimates were made.
‡ Span of region within a radiation hybrid map (<http://www.ncbi.nlm.nih.gov/genome/seq/HsHome.shtml>); 1 Mb $\approx$ 1 centirad. Position of the core SNP is listed first.
§ Number of SNPs discovered with at least 15 copies of the minor allele (successfully genotyped in at least 32 individuals) and in Hardy–Weinberg equilibrium using a significance criterion of $P < 0.02$.
|| The ten regions selected for follow-up genotyping.

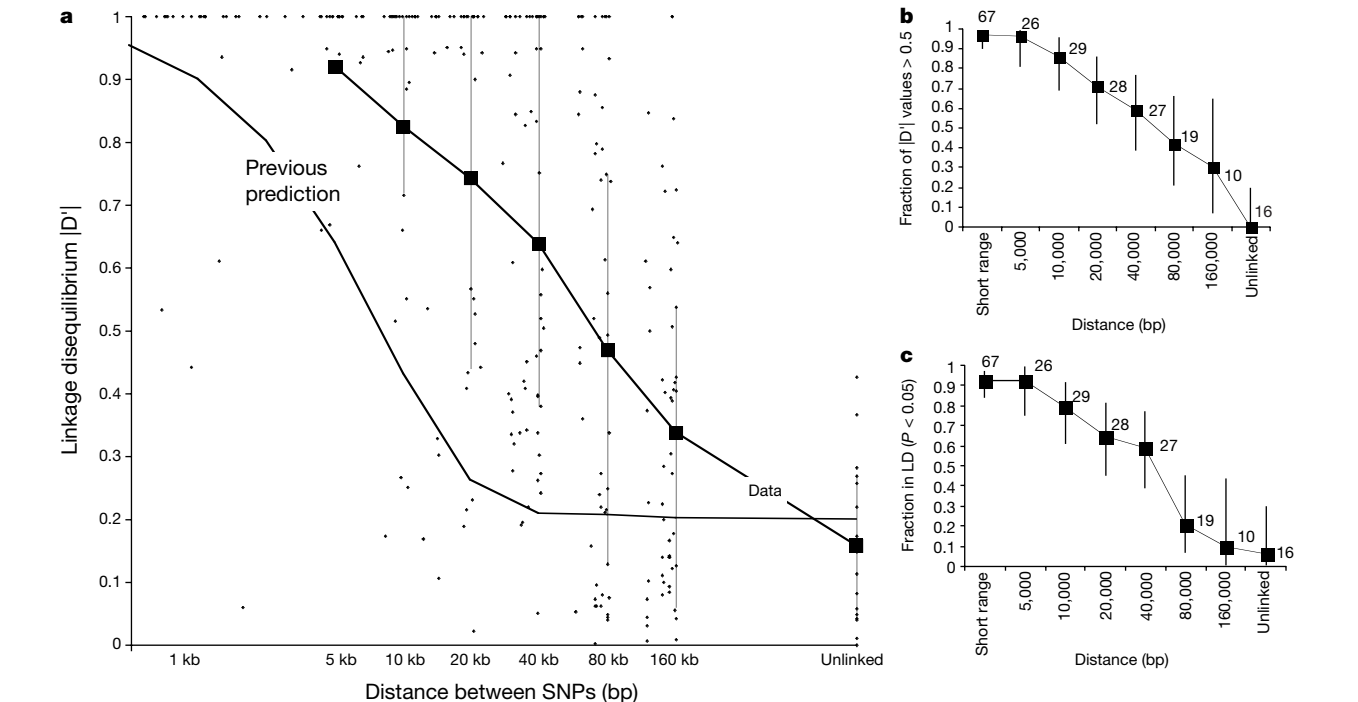


Figure 1 LD versus physical distance between SNPs. For each distance from the core SNP (Table 1), we chose the SNP with the largest number of copies of the minor allele for comparison to SNPs at other distances. At a given distance, all comparisons are independent. **a**, Average $|D'|$ values for each distance separation ('Data'; dotted lines indicate the 25th and 75th percentiles), compared with a prediction² based on simulations (see Methods). $|D'|$ values for shorter physical distances were calculated by looking within contiguously sequenced stretches of DNA containing at least two SNPs, and picking the

two with the most minor alleles. Unlinked marker comparisons are obtained by comparing SNPs in the 40-kb bin in each row of Table 1 to those in the next row. **b**, **c**, Fraction of $|D'|$ values greater than 0.5 (**b**) and proportion of significant ($P < 0.05$) associations (**c**) between two SNPs separated by a given distance (as assessed by a likelihood ratio test¹⁰). Bars indicate 95% central confidence intervals. The number of data points used to make the calculations are shown.

genomic regions (see also ref. 12), which is apparent in the different rates at which LD declines around the core SNP (Fig. 2). For example, $|D'| > 0.5$ for at least 155 kb around the *WASL* gene, but for less than 6 kb around *PCI* (Fig. 2). The variability across different genomic regions within the same population sample provides a context for explaining why past empirical studies, each based on one to three regions^{3–5}, have produced such different results. Large variations in LD are expected because of stochastic factors, such as different gene histories across loci¹³. Differences in recombination rates among regions can also affect the extent of LD. We observe a significant and important correlation ($P < 0.005$) between LD and the estimated local recombination rate (Fig. 2, inset).

Another feature of the data is that, near the range of distances at which LD drops off, there is often considerable variability in $|D'|$

values at neighbouring SNPs (for example, around *IL17R*, *SCYA2*, *TGFB1* and *BMP8* in Fig. 2). Such a wide scatter of LD, even for markers close to each other, has been noted before¹⁴, and is due to the underlying haplotypic structure of LD. SNPs marking sections of chromosome with short extents of correlation are likely to display much lower $|D'|$ values than SNPs marking long haplotypes. In regions of high haplotype diversity, several SNPs may have to be genotyped to have a good chance of tagging most haplotypes. LD-based gene mapping may therefore require clusters of closely spaced SNPs to have maximal power.

Why does LD extend so far? LD around an allele arises because of selection or population history—a small population size, genetic drift or population mixture—and decays owing to recombination, which breaks down ancestral haplotypes¹⁵. The extent of LD

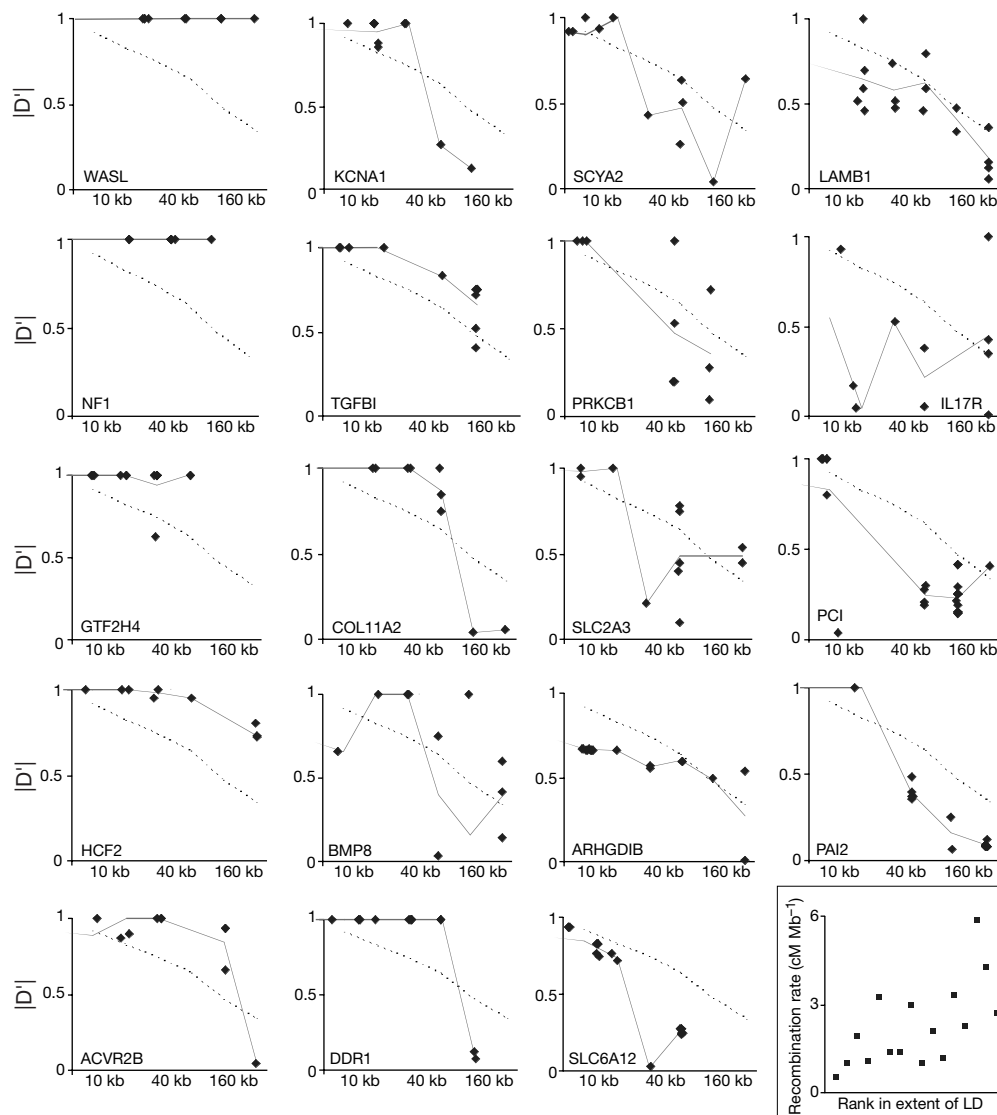


Figure 2 LD profiles for each genomic region. The chosen SNP is usually the core SNP itself, unless the core SNP could not be readily genotyped or another SNP with more minor alleles had been identified within 1 kb. In both cases, we substituted the closest high-frequency SNP. The chosen SNP is compared with every other high-frequency SNP in the same genomic region. Solid line indicates average $|D'|$ values for each distance for which SNPs were available; for comparison, the dashed line indicates the consensus LD curve from Fig. 1. The extent of LD was calculated by performing a least-squares linear regression to the average $|D'|$ values at each distance from the chosen SNP; more sloped lines indicate less LD. The regions are ordered according to the extent of LD (most

extensive LD, top left; least extensive LD, bottom right). Inset (bottom right) shows the rank of each region in terms of LD extent versus the estimated recombination rate per unit of physical distance. For each 160-kb region of interest, we looked for the closest pair of flanking genetic markers from the Marshfield map³⁴ subject to the condition that they were separated by a non-zero genetic distance on the map. We divided genetic map distance by the physical map distance on the basis of the available draft genome sequence³⁵. We analysed the 16 regions for which the genetic and physical map orderings of markers were locally consistent (one-sided Spearman rank correlation, $P < 0.005$).

decreases in proportion to the number of generations since the LD-generating event. The simplest explanation for the observed long-range LD is that the population under study experienced an extreme founder effect or bottleneck: a period when the population was so small that a few ancestral haplotypes gave rise to most of the haplotypes that exist today. Our simulations show that a severe bottleneck (inbreeding coefficient $F \geq 0.2$) occurring 800–1,600 generations ago (about 27,000–53,000 years ago assuming 25 years per generation) could have generated the LD observed (Fig. 3). In principle, long-range LD could also be generated by population mixture¹⁶, but the degree of LD is much greater than would arise from the mixing of even extremely differentiated populations. An alternative explanation for the observed long-range LD is that the recombination rates in the regions studied might be markedly less than the genome-wide average. This could happen if recombination occurred primarily in well-separated hotspots and our regions fell between them (Fig. 3). However, under this hypothesis, the regions would be expected to show long-range LD in all populations, and this pattern is not observed (see below).

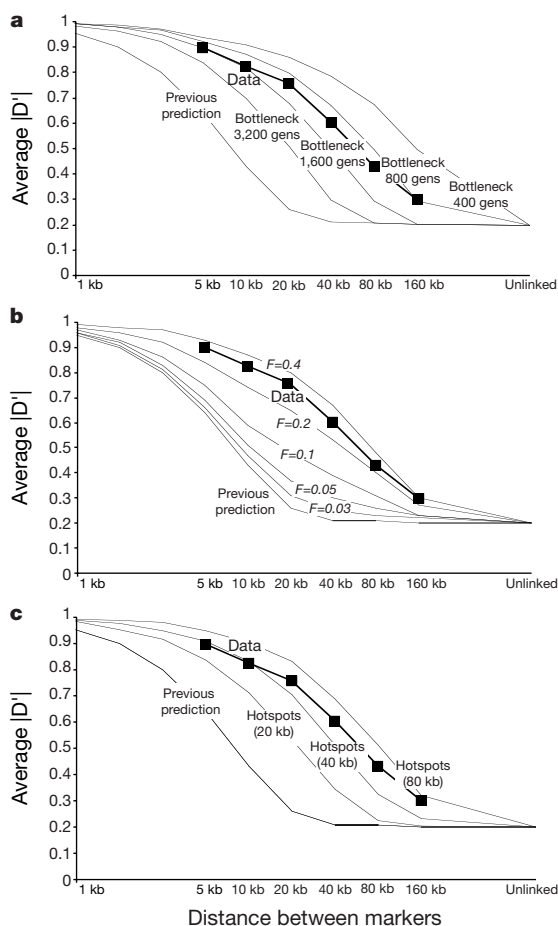


Figure 3 Effect on LD of assumptions about population history and recombination. **a**, The effect of a population bottleneck instantaneously reducing the population to a constant size of 50 individuals for 40 generations ($F = 0.4$) and occurring 400, 800, 1,600 or 3,200 generations ago. **b**, The effect of bottleneck intensity (F) for a bottleneck that occurred 800 generations ago. **c**, The effect of variation in recombination rate, assuming that all recombination in the genome occurs at hotspots randomly distributed according to a Poisson process with an average density of one every 20, 40 and 80 kb, and with a genome-wide average rate of 1.3 centimorgans per megabase per generation³⁵. Results are compared to a no-hotspot model for the same historical hypothesis. ('Data' refers to the mean $|D'|$ values at each physical distance separation, obtained from Fig. 1a).

To confirm our findings of long-range LD and to investigate the reasons for its occurrence, we next examined a representative subset of SNPs in two additional samples. We first studied another north-European sample (48 southern Swedes) and found LD in a nearly identical pattern to that observed in Utah, both in terms of the overall magnitude of LD and the particular alleles that were associated (indicated by the sign of D') (Fig. 4). The similarity in LD patterns may be due to the same historical event, which occurred deep in European prehistory before the separation of the ancestors of these two groups. This suggests that the long-range LD pattern is general in northern Europeans^{3,17}.

We next studied 96 Yorubans (from Nigeria), believed to share common ancestry with northern Europeans about 100,000 years ago¹⁸. At short distances, the Nigerian and European-derived populations typically show the same allelic combinations (Fig. 4): D' has the same sign and a similar magnitude, indicating a common LD-generating event tracing far back in human history. However, the half-length of LD seems to extend less than 5 kb (Fig. 4) in the Yorubans. Markedly shorter range LD in sub-Saharan Africa has also been observed in several studies of single regions^{19,20} (although

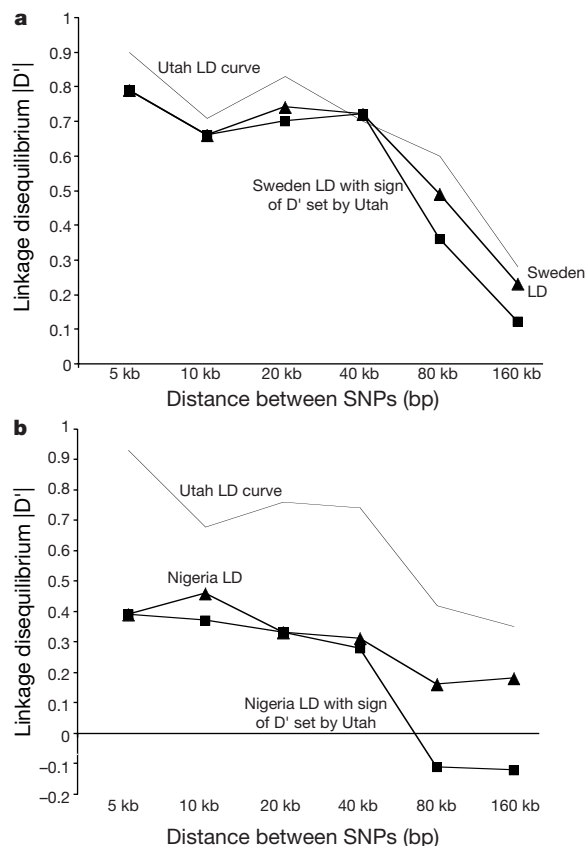


Figure 4 LD curve for Swedish and Yoruban samples. To minimize ascertainment bias, data are only shown for marker comparisons involving the core SNP. Alleles are paired such that $D' > 0$ in the Utah population. $D' > 0$ in the other populations indicates the same direction of allelic association and $D' < 0$ indicates the opposite association. **a**, In Sweden, average D' is nearly identical to the average $|D'|$ values up to 40-kb distances, and the overall curve has a similar shape to that of the Utah population (thin line in **a** and **b**). **b**, LD extends less far in the Yoruban sample, with most of the long-range LD coming from a single region, *HCF2*. Even at 5 kb, the average values of $|D'|$ and D' diverge substantially. To make the comparisons between populations appropriate, the Utah LD curves are calculated solely on the basis of SNPs that had been successfully genotyped and met the minimum frequency criterion in both populations (Swedish and Yoruban).

two other studies did not show a clear trend^{21,22}). Our results indicate that the pattern of shorter LD in sub-Saharan African populations may be general.

Notably, LD in the Nigerians is largely a subset of what is seen in the northern Europeans. The Yoruban haplotypes are generally contained within the longer Utah haplotypes, and there is little Yoruban-specific LD (85% of observations of substantial LD ($|D'| > 0.5$) in Yorubans are also substantial and of the same sign in Utah). The vast difference in the extent of LD between populations points to differences owing to population history, probably a bottleneck or founder effect that occurred among the ancestors of north Europeans after the divergence from the ancestors of the Nigerians. The short extent of LD in Nigerians is more consistent with the predictions of a computer simulation study assuming a simple model of population expansion².

Could the apparent differences in the extent of LD among populations be due to 'ascertainment bias' in the identification of the SNPs? The core SNPs are probably not subject to bias because they were identified in a multi-ethnic population. The neighbouring SNPs were identified in the Utah population and subsequently studied in the other populations, and thus they may be susceptible to ascertainment bias. However, we selected only SNPs with high frequency in Utah and most of these satisfied the high-frequency criterion for use in the other populations (87% in Sweden and 71% in Nigeria). Thus, the inferences about LD are not likely to be much different from what would have been obtained had we used SNPs ascertained in the Yoruban sample. Moreover, the cross-population comparisons (Fig. 4) minimize ascertainment bias because they involve only the core SNP, and because they calculate LD in each population using only the SNPs present in both.

What was the nature of the population event that created the long-range LD? The event could be specific to northern Europe, which was substantially depopulated during the Last Glacial Maximum (30,000–15,000 years ago), and subsequently recolonized by a small number of founders^{23,24}. Alternatively, the long-range LD could be due to a severe bottleneck that occurred during the founding of Europe or during the dispersal of anatomically modern humans from Africa^{19,20,25,26} (the proposed 'Out of Africa' event) as recently as 50,000 years ago. Under the first hypothesis, the strong LD at distances ≥ 40 kb would be absent in populations not descended from northern Europeans. Under the second hypothesis, the same pattern of long-range LD could be observed in a variety of non-African populations. Regardless of the timing and context of the bottleneck, the severity of the event (in terms of inbreeding) can be assessed from our data. To have a strong effect on LD, a substantial proportion of the modern population would have to be derived from a population that had experienced an event leading to an inbreeding coefficient of at least $F = 0.2$ (Fig. 3). This corresponds to an effective population size (typically less than the true population size¹⁵) of 50 individuals for 20 generations; 1,000 individuals for 400 generations; or any other combination with the same ratio.

Our results have implications for disease gene mapping, suggesting a possible two-tiered strategy for using LD. The presence of large blocks of LD in northern European populations suggests that genome-wide LD mapping is likely to be possible with existing SNP resources¹. Although the large blocks should make initial localization easier, they may also limit the resolution of mapping to blocks of DNA in the range of 100 kb²⁷. Populations with much smaller blocks of LD (for example, Yorubans) may allow fine-structure mapping to identify the specific nucleotide substitution responsible for a phenotype¹². Our study also has implications for LD as a tool to study population history^{19–22}. Simultaneous assessment of LD at multiple regions of the genome provides an approach for studying history with potentially greater sensitivity to certain aspects of history than traditional methods based on properties of a single locus. □

Methods

Core SNPs were identified by screening more than 3,000 genes in a multi-ethnic panel of 15 European Americans, 10 African Americans, and 7 East Asians (see ref. 28 for details; a full description of this database will be presented elsewhere). DNA used for sequencing was obtained from the Coriell Cell Repositories. Identification numbers for these Utah samples from the CEPH mapping panel were NA12344, 06995, 06997, 07013, 12335, 06990, 10848, 07038, 06987, 10846, 10847, 07029, 07019, 07048, 06991, 10851, 07349, 07348, 10857, 10852, 10858, 10859, 10854, 10856, 10855, 12386, 12456, 10860, 10861, 10863, 10830, 10831, 10835, 10834, 10837, 10836, 10838, 10839, 10841, 10840, 10842, 10843, 10845 and 10844. We did follow-up genotyping in 48 Swedes (healthy individuals from a case/control study of adult-onset diabetes) and in 96 Yoruban males from Nigeria (healthy individuals from a case/control study of hypertension).

SNPs were discovered by DNA sequencing²⁸ in the 44 individuals from Utah; we sequenced about 2 kb centred at each distance from the core SNP ≥ 5 kb, with about 1 kb sequenced around the core SNP itself. When no polymorphism of sufficiently high frequency was found, a nearby subregion of about two further kilobases was resequenced; this occurred in only 18% of the cases. Polymorphisms were identified and genotypes were scored automatically using Polyphred²⁹ and checked manually by at least two different scorers. SNPs in Hardy–Weinberg disequilibrium or showing evidence of breakdown of LD over short physical distances (< 2.5 kb) were triple-checked. Of the 275 high-frequency SNPs (that is, SNPs with at least 15 observed copies of the minor allele), three were discarded because of a Hardy–Weinberg P value of < 0.02 ; one of the SNPs used in the analysis had a nominally significant P value ($P < 0.05$). To assess the accuracy of scoring, we rescored 26 randomly chosen high-frequency SNPs; only seven discrepancies were found among 1,144 genotypes. For cases in which follow-up genotyping was done, the discrepancy rate was 47 out of 1,484 (3%) between genotypes obtained by both methods.

Genotyping of SNPs was performed by single-base extension followed by mass spectroscopy (Sequenom)³⁰, fluorescence polarization (LJL Biosystems)³¹ or detection on a sequencing gel (Applied Biosystems)³². For the ten regions selected for follow-up genotyping (Table 1), we chose at most one 'representative' SNP at each distance from the core SNP (each column in Table 1) according to the criterion that it had the highest number of minor alleles of all SNPs at that distance from the core SNP. For other populations, only those SNPs that, when genotyped, had a minimum number of minor alleles were included in studies of LD. For Yorubans, the cutoff was 25 alleles (76% of SNPs met the criterion); for Swedes, the cutoff was 15 alleles (89%). The fact that most of the SNPs we studied in Utah are also present in high frequency in these other populations indicates that the assessment of LD is not likely to be subject to large ascertainment bias.

Heterozygosity¹⁵ (π) was calculated as the average of $2jk/n(n-1)$ for all base pairs screened, with j and k equal to the number of copies of the minor and major alleles, respectively ($n = j + k$). A base was considered screened if it had Phred quality scores²⁹ of ≥ 15 in ≥ 10 individuals. D' values between markers with alleles A/a and B/b (allele frequencies, c_A , c_a , c_B and c_b ; haplotype frequencies, c_{AB} , c_{Ab} , c_{aB} and c_{ab}) were obtained by dividing $c_{AB} - c_Ac_B$ by its maximum possible value: $\min(c_Ac_b, c_a c_B)$ if $D > 0$ and $\min(c_Ac_b, c_a c_B)$ otherwise. An implementation of the expectation maximization algorithm was used to infer haplotype frequencies for pairs of SNPs both for actual and simulated data³³. A likelihood ratio test was used to assess significance of associations between pairs of SNPs¹⁰.

Computer simulations were based on a model related to that in ref. 2, assuming a population that was constant at an effective size of 10,000 individuals until 5,000 generations ago, when it expanded instantaneously to a size of 100,000,000 (an arbitrary value). This model captures many of the features of more complicated growth, as the effect of population growth on LD is not dependent on the precise details of population growth or the final population size when the growth is moderately fast². Bottlenecks were modelled as described, with the population crashing to a constant size for a fixed number of generations before re-expanding. (The effect of a bottleneck on LD depends primarily on the F -value, the inbreeding coefficient, which is defined as the probability that two alleles randomly picked from the population after the bottleneck derive from the same ancestral allele just before the bottleneck.) Coalescent simulations were used to generate gene genealogies under these models for markers separated by a specified recombination distance (see ref. 13 for a more detailed description of the theory behind these simulations). Simulations were run 2,000 times with sample-size distributions mimicking our data. SNPs were generated by distributing mutations on the simulated gene genealogies at a mutation rate of 6×10^{-5} per generation, under an 'infinite alleles' mutation model. The mutation rate was chosen such that the probability of high frequency SNPs in a 2-kb stretch of DNA sequenced in 44 samples (for the model of a simple expansion 5,000 generations ago) was similar to what we observed (about 70%). We also tested mutation rates ten times higher and found that inferences about LD were essentially unchanged.

An extreme hypothesis of population mixture and its effect on LD were assessed in a simulated, mixed population of European Americans and sub-Saharan Africans. For the first simulated mixture, we constructed a mixture of 22 Yorubans and 22 samples from Utah, and used data from the ten core SNPs genotyped in both populations. For the second simulation, we used 26 SNPs from a previous study⁷ that had been found to have a minor allele frequency of at least 15% in either African Americans or European Americans; we chose at most one SNP per gene, picking the first listed SNP (in Table A1 of ref. 7) that met our minimum frequency criterion. We found much stronger LD even at 40 kb (56% with $|D'| > 0.5$) (Fig. 1) in our actual data than in the simulated, admixed populations. For the 45 possible pairwise comparisons of the ten core SNPs in the simulated mix of Yoruban and Utah samples, no values of $|D'| > 0.5$ were observed. For the 325 possible

pairwise comparisons from the second study, only 11% showed $|D'| > 0.5$. This suggests that admixture probably did not generate the strong signal of LD at long physical distances seen in Utah.

Received 11 December 2000; accepted 13 March 2001.

- Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
- Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
- Dunning, A. M. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67**, 1544–1554 (2000).
- Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
- Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
- Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
- Goddard, K. A. B., Hopkins, P. J., Hall, J. M. & Witte, J. S. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am. J. Hum. Genet.* **66**, 216–234 (2000).
- Watterson, G. A. & Guess, H. A. Is the most frequent allele the oldest? *Theor. Pop. Biol.* **11**, 141–160 (1977).
- Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
- Schneider, S., Kueffler, J. M., Roessli, D. & Excoffier, L. Arlequin (ver. 2.0): A software for population genetic data analysis (Genetics and Biometry Laboratory, Univ. Geneva, Switzerland, 2000).
- Lewontin, R. C. On measures of gametic disequilibrium. *Genetics* **120**, 849–852 (1988).
- Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
- Hudson, R. R. in *Oxford Surveys in Evolutionary Biology* (eds Futuyma, D. J. & Antonovics, J.) 1–44 (Oxford Univ. Press, Oxford, 1990).
- Clark, A. G. *et al.* Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* **63**, 595–612 (1998).
- Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* (Sinauer, Massachusetts, 1997).
- Chakraborty, R. & Weiss, K. M. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl Acad. Sci. USA* **85**, 9119–9123 (1988).
- Eaves, I. A. *et al.* The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nature Genet.* **25**, 320–322 (2000).
- Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl Acad. Sci. USA* **92**, 6723–6727 (1995).
- Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
- Tishkoff, S. A. *et al.* Short tandem-repeat polymorphism/Alu haplotype variation at the PLAT locus: Implications for modern human origins. *Am. J. Hum. Genet.* **67**, 901–925 (2000).
- Kidd, J. R. *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am. J. Hum. Genet.* **66**, 1882–1899 (2000).
- Mateu, E. *et al.* Worldwide genetic analysis of the CFTR region. *Am. J. Hum. Genet.* **68**, 103–117 (2001).
- Housley, R. A., Gamble, C. S., Street, M. & Pettitt, P. Radiocarbon evidence for the Late glacial human recolonisation of northern Europe. *Proc. Prehist. Soc.* **63**, 25–54 (1994).
- Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000).
- Reich, D. E. & Goldstein, D. B. Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl Acad. Sci. USA* **95**, 8119–8123 (1998).
- Ingman, M., Kaessmann, H., Pääbo, S. & Gyllenstein, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
- Altshuler, D., Daly, M. & Kruglyak, L. Guilt by association. *Nature Genet.* **26**, 135–137 (2000).
- Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
- Nickerson, D. B., Tobe, V. O. & Taylor, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based sequencing. *Nucleic Acids Res.* **25**, 2745–2751 (1997).
- Ross, P. Hall, L., Smirnov, I. & Haff, L. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nature Biotech.* **16**, 1347–1351 (1998).
- Chen, X., Levine, L. & Kwok, P. Y. Fluorescence polarization in homogenous nucleic acid analysis. *Genome Res.* **9**, 492–498 (1999).
- Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
- Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**, 921–927 (1995).
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–889 (1998).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Acknowledgements

We thank L. Groop for the Swedish samples; R. Cooper and C. Rotimi for the Yoruban samples; and D. Altshuler, M. Daly, D. Goldstein, J. Hirschhorn, C. Lindgren and S. Schaffner for discussions. This work was supported in part by grants from the National Institutes of Health, Affymetrix, Millennium Pharmaceuticals, and Bristol-Myers Squibb Company, and by a National Defense Science and Engineering Fellowship to D.E.R.

Correspondence and requests for materials should be addressed to D.E.R. (e-mail: reich@genome.wi.mit.edu) or E.S.L. (e-mail: lander@genome.wi.mit.edu).

The *Ph1* locus is needed to ensure specific somatic and meiotic centromere association

Enrique Martinez-Perez, Peter Shaw & Graham Moore

John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK

The correct pairing and segregation of chromosomes during meiosis is essential for genetic stability and subsequent fertility. This is more difficult to achieve in polyploid species, such as wheat, because they possess more than one diploid set of similar chromosomes. In wheat, the *Ph1* locus ensures correct homologue pairing and recombination¹. Although clustering of telomeres into a bouquet early in meiosis has been suggested to facilitate homologue pairing^{2,3}, centromeres associate in pairs in polyploid cereals early during floral development⁴. We can now extend this observation to root development. Here we show that the *Ph1* locus acts both meiotically and somatically by reducing non-homologous centromere associations. This has the effect of promoting true homologous association when centromeres are induced to associate. In fact, non-homologously associated centromeres separate at the beginning of meiosis in the presence, but not the absence, of *Ph1*. This permits the correction of homologue association during the telomere-bouquet stage in meiosis. We conclude that the *Ph1* locus is not responsible for the induction of centromere association, but rather for its specificity.

We previously showed that centromeres associate in pairs before meiosis in polyploid cereals, but not until the beginning of meiosis in their diploid progenitors⁴. Using fluorescence *in situ* hybridization on intact root sections, we now report that centromeres also associate in pairs in developing xylem vessel cells of bread wheat (AABBDD, $2n = 6x = 42$) but not in those of its diploid progenitors (Fig. 1b, e, and Table 1). Moreover, we show that during this developmental process the chromosomes endoreplicate, becoming polytene. This is indicated by the substantial increase in size of the interphase chromosomes (and nucleus), as compared with the surrounding tissues (Fig. 1a, d, f, g).

The level of centromere association in xylem vessel cells of wheat is unaffected by the presence of *Ph1*, as in floral development⁵ (Fig. 1e, h). Thus, neither endoreplication nor the *Ph1* locus can induce centromere association. Although centromeres associate in the xylem vessel cells in the presence and absence of *Ph1*, they are not associated in other root tissues (Fig. 1c). Polyploidy is therefore necessary but not sufficient to induce centromere association—a specific developmental context is also required, as in meiosis, floral development or xylem vessel development.

We have assessed homologue association in these vessel cells by labelling specific pairs of rye chromosomes in wheat–rye addition lines. These rye homologues associate at a high level through their centromeres during vessel development in the presence of *Ph1* (25/25

Table 1 Statistics of the number of centromeres

		<i>Ph1</i> +	<i>Ph1</i> –	<i>t</i> -test
Wheat–rye	Premeliosis	19.5 (2.7)	16.3 (2.9)	$P < 0.001$
	Telomere bouquet	23.5 (1.5)	13.7 (2.1)	$P < 0.001$
	Xylem vessel	20 (1.7)	16.3 (1.8)	$P < 0.001$
	Non-polytene root	24.9 (1.2)	25.2 (1.3)	$P = 0.6$
Wheat	Xylem vessel	21.7 (2.1)	21.5 (1.8)	$P = 0.798$
<i>T. monococtum</i>	Xylem vessel	12.56 (0.8)		

The s.d. is given in parentheses. Student's *t*-test was used to test the null hypothesis that the two means in the presence and absence of *Ph1* are the same. The null hypothesis can be discounted in all comparisons except the wheat xylem and the wheat–rye non-polytene root. All centromere sites were counted on the original three-dimensional confocal stacks.

cells at the last stage of development examined) but not in its absence (22/25 cells scored did not show association) (Fig. 1f, g, i, j). In both the presence and absence of *Ph1*, as we have shown above, the number of centromere sites seen indicates centromere pairing (Fig. 1e, h). This suggests that in the presence of *Ph1* most centromeres are homologously associated. In other root cells, neither the centromeres nor the homologues are associated (Fig. 1c, d). In xylem cells in which the homologues are associated at their centromeres, associations also occur at interstitial sites in 67% of the cells. But no interstitial associations are observed unless the homologue centromeres are also associated, indicating that centromere association may precede interstitial association.

Homologous chromosomes also associate somatically in many different *Drosophila* tissues^{6,7}, and numerous studies have shown that this association induces an extra level of epigenetic regulation through trans-sensing or transvection⁶⁻⁸. In common with *Drosophila*, wheat chromosomes are organized in a Rabl configu-

ration with the centromeres grouped at one pole of the nucleus and the telomeres at the opposite pole^{9,10}. In this configuration, association of homologous centromeres will facilitate the interaction of other chromosomal regions in the large wheat nucleus. Thus, the presence of *Ph1* in wheat has the potential to induce epigenetic control, in the same manner as in *Drosophila*. This level of epigenetic regulation is not available to diploid cereals, and may be a factor in the widespread occurrence of polyploidy in the plant kingdom.

The effect of *Ph1* in the xylem vessel cells is analogous to the premeiotic association observed during floral development¹¹. Thus, in those somatic tissues in which the centromeres associate, *Ph1* would promote the specificity of the association to true homologues, and this specificity could be achieved by reducing non-homologous centromere associations. To test this hypothesis, we determined the centromere behaviour during meiosis, floral development and xylem vessel development in hybrid lines possessing only non-homologous chromosomes, in both the presence and

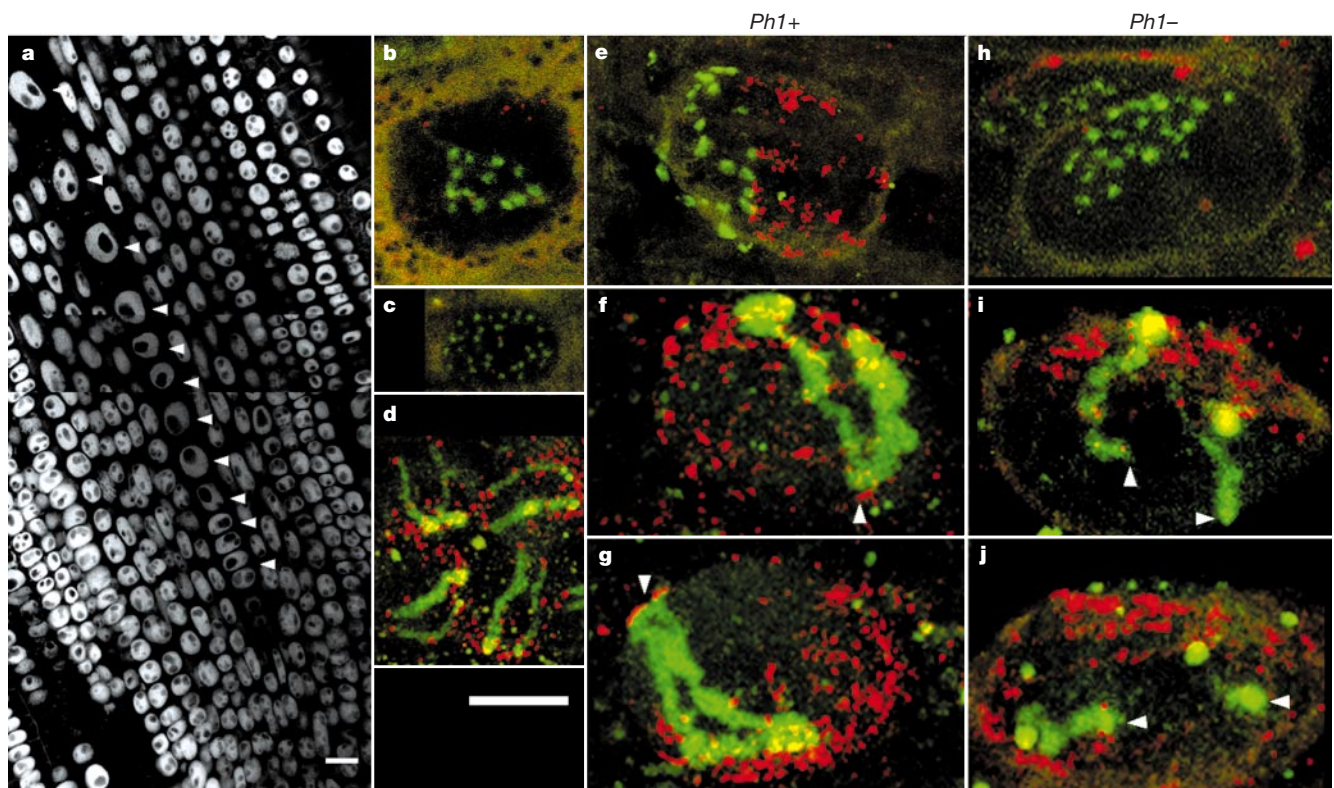


Figure 1 Centromere and homologue behaviour during root development. **a**, Root section labelled with DAPI (4',6-diamidino-2-phenylindole dihydrochloride). Arrowheads indicate the column of developing xylem vessel cells. **b**, Polytene vessel nucleus in *T. monococcum*, $2n = 14$, showing 13 centromeric sites (green). **c**, Non-polytene root nucleus in hexaploid wheat, $2n = 42$, showing 35 centromeric sites (green). **d**, Non-polytene root nuclei in hexaploid wheat, showing the homologues unassociated (green).

e, h, Polytene vessel nuclei in hexaploid wheat, showing about 21 centromere sites (green) in both the presence (**e**) and absence (**h**) of *Ph1*. **f, g**, Polytene vessel nuclei in hexaploid wheat, showing homologues (green) associated through their centromeres (arrowheads). **i, j**, Polytene vessel nuclei (*Ph1* mutant), showing homologues (green) unassociated; centromeres are indicated by arrowheads. Telomeres are red in **b–j**, which are all at same magnification. Scale bars, 10 μm .

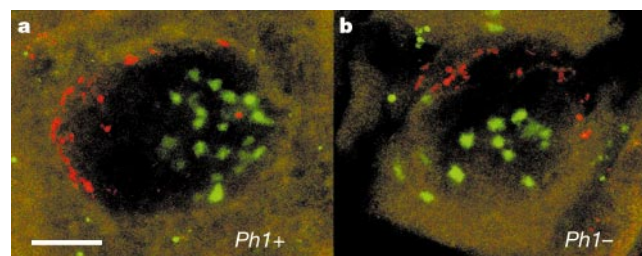


Figure 2 Centromeres (green) and telomeres (red) in root xylem vessel cells of wheat-rye hybrids. **a**, Presence of *Ph1*, showing 20 centromere sites. **b**, Absence of *Ph1*, showing 14 centromere sites. Scale bar, 10 μm .

absence of the *Ph1* locus. These lines are wheat-rye hybrids, and contain a haploid set of 21 wheat chromosomes and a haploid set of 7 rye chromosomes, making 28 chromosomes in total.

In most root tissues, centromeres are not associated (Table 1). During xylem vessel development in the absence of the *Ph1* locus, however, the centromeres associate, reducing to a mean of 16 centromere sites, whereas in the presence of the *Ph1* locus, a mean of 20 sites is seen (Table 1, and Fig. 2a, b). Similarly, during floral development, a mean of 16 paired centromere sites is formed in the absence of the *Ph1* locus and 19 in its presence (Table 1, and Fig. 3). Thus, in these somatic cells, *Ph1* can reduce but not eliminate non-homologous association. The seven rye chromosomes are visualized as up to seven domains in both the presence and absence of *Ph1*, showing that there is no physical separation of the genomes and that wheat and rye centromeres associate (Fig. 3c, h).

The telomere bouquet is formed in these hybrids at the onset of meiosis in both the presence and absence of *Ph1* (Fig. 3k, l). At the

telomere-bouquet stage, in the absence of the *Ph1* locus the centromeres reduce to a mean of 14 sites, showing complete but non-homologous pairing. In the presence of *Ph1* the mean number of sites increases to 24 (Table 1, and Fig. 3l). Previous studies have indicated that in these lines and in other hybrids in which only non-homologous chromosomes are present, the telomere and interstitial regions can synapse to the same degree in both the presence and absence of *Ph1* (refs 12, 13).

This confirms that, at the time of synapsis, the *Ph1* locus increases the specificity of interaction between true homologues not at these sites, but rather at the centromeres, as we have shown. Non-homologous centromere associations are mostly eliminated at the telomere-bouquet stage of meiosis in wheat-rye hybrids in the presence of *Ph1*, but not in its absence. Because higher levels of recombination occur in the absence of *Ph1* than in its presence^{14,15}, we suggest that the maintenance of centromere association is important for recombination to occur at all in these hybrids.

We can now explain the previous studies on chromosome synapsis in the presence and absence of *Ph1* (refs 16–18). In tetraploid and hexaploid wheat, the centromeres associate during floral development whether *Ph1* is present or not; however, a higher proportion of paired sites will be homologous in the presence of *Ph1*. This premeiotic association of centromeres provides an initial sorting of the chromosomes. The sites remain paired through floral development to the telomere-bouquet stage of meiosis. At the telomere-bouquet stage, any non-homologous centromere associations separate in the presence of *Ph1*, enabling a further sorting of these chromosomes.

Studies of synapsis in the presence of *Ph1* indicate that most chromosomes synapse as bivalents, although some multivalents are formed early on (involving on average a total of six chromosomes)¹⁶. As the synapsis of chromosomes progresses the number of multivalents declines, suggesting a correction of pairing, in agreement with our observations. In the absence of *Ph1*, non-homologous associations made before meiosis are maintained. Thus, telomere-led synapsis results in a high level of multivalents (involving a total of up to 17 chromosomes), which remain uncorrected^{17,18}. As a result, at least 14 chromosomes are incorrectly paired at metaphase I (ref. 19).

Deletion of the *Ph1* locus has two major phenotypic effects: first, chromosome pairing is disrupted, which leads to the synapsis of non-homologous chromosomes; second, non-homologous recombination is induced²⁰. The effect on chromosome pairing has been proposed to occur either premeiotically²¹ or during synapsis^{17,18}. Our data suggest that the effect of *Ph1* on premeiotic alignment and correction during synapsis can be explained by a mechanism involving centromere association. At meiosis, any pre-existing incorrect associations will disrupt chromosome pairing unless separated and re-paired correctly.

This might be achieved either by the separation of all centromere associations or by a more targeted separation solely of incorrect associations. Our studies suggest that the latter occurs. Linking two non-homologous chromosome arms through a single centromere is not sufficient to induce recombination between them in the presence of *Ph1*, showing that the *Ph1* recombination phenotype cannot be simply explained by centromere association²². This phenotypic effect may be the indirect consequence of globally disrupting chromosome pairing, which results in the induction of non-homologous recombination²³. As yet, however, we cannot exclude the possibility that the *Ph1* locus contains more than one gene with individual phenotypic effects.

The mechanism that we have shown for the effect of *Ph1* on homologue pairing depends on the presence of chromosome-specific centromeres. It has been previously shown in cereals that centromeres can evolve rapidly leading to chromosome-specific organization^{24,25}. Our results reveal a function that exploits the specificity of these macromolecular structures. □

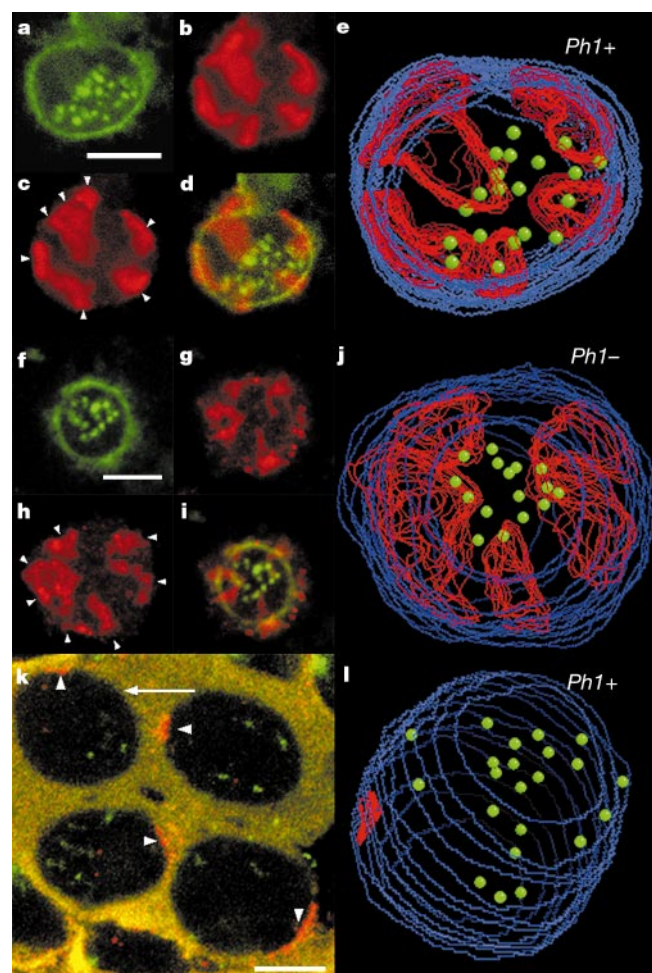


Figure 3 Premeiotic and meiotic behaviour of centromeres in wheat-rye hybrids. **a–e**, Premeiotic pollen mother cell (*Ph1* present). **a**, Single confocal section showing 20 centromere sites. **b**, equivalent section to **a** showing five rye chromosome domains. **c**, Confocal section showing seven rye chromosomal domains (arrowheads). **d**, Overlay of **a** and **b**. **e**, Three-dimensional reconstruction of the nucleus shown in **a–d**. **f–j**, Premeiotic pollen mother cell (*Ph1* absent). **f**, Single confocal section showing 15 centromere sites out of the 16 present in the entire three-dimensional stack. **g**, Equivalent section to **f**, showing four rye chromosome domains. **h**, Confocal section showing seven rye chromosome domains (arrowheads). **i**, Overlay of **f** and **g**. **j**, Three-dimensional model of the nucleus shown in **f–i**. **k**, Confocal section showing four meiocytes possessing telomere bouquets (red, arrowheads) in the presence of *Ph1*, centromeres shown in green. **l**, Three-dimensional model of the nucleus arrowed in **k**, showing 22 centromere sites. Scale bars, 10 μ m.

Methods

Plant material

The roots and anthers used in this study came from hexaploid wheat (*Triticum aestivum*, cv. Chinese spring); a Chinese spring mutant (*ph1b*) lacking the *Ph1* locus; a *Ph1* mutant line carrying 1RS chromosome arm substitution for 1BS wheat chromosome arm; a hexaploid wheat addition line carrying an extra pair of 1RL (*Secale cereale*) telocentric chromosomes; Chinese spring/*Secale cereale* cv. Petkus F₁ hybrids with and without the *Ph1* locus; and the diploid wheat progenitor, *Triticum monococcum*. All the lines lacking *Ph1* used here carried the *ph1b* deficiency.

Sectioning and fluorescence *in situ* hybridization

The centromere (CCS1) and telomere (TTTAGGG repeat) probes used, tissue sectioning and specimen preparation, *in situ* hybridization and probe preparation, and the labelling of the two homologous rye arms (see Fig. 1f, g, i, j) have all been described^{4,5}.

Fluorescence microscopy and image processing

We collected confocal optical section stacks using a Leica TCS SP or SP2 confocal microscope as described⁵. Confocal images were processed by Adobe Photoshop and the public domain program NIH Image (W. Rasband). The models were created using Object-image 1.62n7 (a modified version of NIH Image; N. O. Visscher) and Rotater 3.5.

Received 24 November 2000; accepted 1 March 2001.

- Riley, R. & Chapman, V. Genetic control of cytologically diploid behaviour of hexaploid wheat. *Nature* **182**, 713–715 (1958).
- Rockmill, B. & Roeder, G. S. Telomere-mediated chromosome pairing during meiosis in budding yeast. *Genes Dev.* **12**, 2574–2586 (1998).
- Trelles-Sticken, E., Dresser, M. E. & Scherthan, H. Meiotic Telomere protein Ndj1p is required for meiosis-specific telomere distribution, bouquet formation and efficient homologue pairing. *J. Cell Biol.* **151**, 95–106 (2000).
- Martinez-Perez, E., Shaw, P. & Moore, G. Polyploidy induces centromere association. *J. Cell Biol.* **148**, 233–238 (2000).
- Martinez-Perez, E. *et al.* Homologous chromosome pairing in wheat. *J. Cell Sci.* **112**, 1761–1769 (1999).
- Henikoff, S. Nuclear organisation and gene expression: homologous pairing and long-range interactions. *Curr. Opin. Cell Biol.* **9**, 388–395 (1997).
- Gemkov, M. J., Vermeer, P. J. & Arndt-Jovin, D. J. Homologous association of Bitorax-Complex during embryogenesis: consequences for transvection in *Drosophila melanogaster*. *Development* **125**, 4541–4552 (1998).
- Wolfe, A. M. & Matzke, M. A. Epigenetics: regulation through repression. *Science* **286**, 481–486 (1999).
- Marshall, W. F. *et al.* Specific interactions of chromatin with the nuclear envelope: positional determination within the nucleus in *Drosophila melanogaster*. *Mol. Biol. Cell* **7**, 825–842 (1996).
- Abranches, R., Beven, A. F., Aragon-Alcaide, L. & Shaw, P. Transcription sites are not correlated with chromosome domains in wheat nuclei. *J. Cell Biol.* **143**, 5–12 (1998).
- Aragon-Alcaide, L. *et al.* Association of homologous chromosomes during floral development. *Curr. Biol.* **7**, 905–908 (1997).
- Wang, X. & Holm, P. B. Chromosome pairing and synaptonemal complex formation in wheat × rye hybrids. *Carlsberg Res. Commun.* **53**, 167–190 (1988).
- Gillies, C. B. The effect of the *Ph1* gene alleles on synaptonemal complex formation in *Triticum aestivum* × *T. kotschy* hybrids. *Theor. Appl. Genet.* **74**, 430–438 (1987).
- Dhalwal, H. S., Gill, B. S. & Waines, J. G. Analysis of induced homoeologue pairing I the *ph1* mutant wheat × rye hybrid. *J. Hered.* **68**, 206–209 (1977).
- Dvorak, J. Effect of rye on homoeologous chromosome pairing in wheat × rye hybrids. *Can. J. Genet. Cytol.* **19**, 549–556 (1977).
- Holm, P. B. Chromosome pairing and chiasma formation in allohexaploid wheat, *Triticum aestivum* analyzed by spreading of meiotic nuclei. *Carlsberg Res. Commun.* **51**, 239–294 (1986).
- Holm, P. B. Chromosome pairing and synaptonemal complex formation in hexaploid wheat nullisomic for chromosome 5B. *Carlsberg Res. Commun.* **53**, 91–110 (1988).
- Holm, P. B. & Wang, X. The effect of chromosome 5B on synapsis and chiasma formation in wheat, *Triticum aestivum* cv Chinese Spring. *Carlsberg Res. Commun.* **53**, 191–208 (1988).
- Roberts, M. A. *et al.* Induction and characterisation of *Ph1* wheat mutants. *Genetics* **153**, 1909–1918 (1999).
- Wall, A. M., Riley, R. & Gale, M. D. The position of a locus on chromosome 5B of *Triticum aestivum* affecting homoeologous meiotic pairing. *Genet. Res.* **18**, 329–339 (1971).
- Feldman, M. The effect of chromosome 5A, 5B and 5D on chromosome pairing in *Triticum aestivum*. *Proc. Natl Acad. Sci. USA* **55**, 1447–1453 (1966).
- Dvorak, J. & Lukaszweski, A. J. Centromere association is an unlikely mechanism by which the wheat *Ph1* locus regulates metaphase I chromosome pairing between homoeologous chromosomes. *Chromosoma* **109**, 410–414 (2000).
- Riley, R. & Miller, T. E. Meiotic chromosome pairing in triticales. *Nature* **227**, 82–83 (1970).
- Langdon, T. *et al.* Retrotransposon evolution in diverse plant genomes. *Genetics* **156**, 313–325 (2000).
- Ananiev, E. V., Phillips, R. L. & Rines, H. W. Chromosome specific molecular organisation of maize (*Zea mays* L.) centromeric regions. *Proc. Natl Acad. Sci. USA* **95**, 13073–13078 (1998).

Acknowledgements

This work was funded by the Biotechnology and Biological Sciences Research Council of the UK, and the John Innes foundation.

Correspondence and requests for materials should be addressed to G.M. (e-mail: graham.moore@bbsrc.ac.uk).

Phagocytosis and clearance of apoptotic cells is mediated by MER

Rona S. Scott^{*†}, Eileen J. McMahon[‡], Shannon M. Pop[§], Elizabeth A. Reap^{||}, Roberto Caricchio[¶], Philip L. Cohen[¶], H. Shelton Earp[#] & Glenn K. Matsushima^{*‡§☆}

^{*} UNC Neuroscience Center, [‡] Department of Microbiology and Immunology, [§] Curriculum in Oral Biology, [#] Department of Pharmacology and Lineberger Comprehensive Cancer Center, [☆] Comprehensive Center for Inflammatory Disorders and Program for Molecular Biology and Biotechnology, University of North Carolina at Chapel Hill, North Carolina 27599, USA ^{||} Alphavax Human Vaccines, Durham, North Carolina 27701, USA [¶] University of Pennsylvania, Department of Medicine, Philadelphia, Pennsylvania 19104, USA

[†] Present address: Louisiana State University, Health Sciences Center, Shreveport, Louisiana, USA

Apoptosis is fundamental to the development and maintenance of animal tissues and the immune system¹. Rapid clearance of apoptotic cells by macrophages is important to inhibit inflammation and autoimmune responses against intracellular antigens^{2–4}. Here we report a new function for Mer, a member of the Axl/Mer/Tyro3 receptor tyrosine kinase family. *mer^{kd}* mice with a cytoplasmic truncation of Mer had macrophages deficient in the clearance of apoptotic thymocytes. This was corrected in chimaeric mice reconstituted with bone marrow from wild-type animals. Primary macrophages isolated from *mer^{kd}* mice showed that the phagocytic deficiency was restricted to apoptotic cells and was independent of Fc receptor-mediated phagocytosis or ingestion of other particles. The inability to clear apoptotic cells adequately may be linked to an increased number of nuclear autoantibodies in *mer^{kd}* mice. Thus, the Mer receptor tyrosine kinase seems to be critical for the engulfment and efficient clearance of apoptotic cells. This has implications for inflammation and autoimmune diseases such as systemic lupus erythematosus.

Several receptors have been identified that participate in the recognition and phagocytosis of apoptotic cells. These include the scavenger receptor class A, CD36 and the phosphatidylserine receptor, which bind phosphatidylserine on the surface of apoptotic cells; the vitronectin receptor (αvβ3 integrin) in cooperation with CD36, which use thrombospondin as a molecular bridge to recognize apoptotic cells; and CD14, the lipopolysaccharide (LPS) receptor, and complement, which recognize an unknown ligand on apoptotic cells^{5–12}. The family of Axl/Mer/Tyro3 receptor tyrosine kinases have several functions in different tissues^{13–17}. Previously, we showed that a functional knockout mouse with a truncated cytoplasmic tail of the Mer receptor tyrosine kinase (*mer^{kd}*) was hypersensitive to high doses of LPS, and suggested that Mer may

Table 1 Flow cytometry analysis of the percentage of dead cells present in the thymus of dexamethasone-treated mice

Group	Mean (% dead cells)	Fold increase compared with WT
Wild type	4.40	1
<i>mer^{kd}</i>	32.38	7.4
Saline controls	1.05	–
Bone marrow chimaeras*		
WT into WT	1.10	1
<i>mer^{kd}</i> into <i>mer^{kd}</i>	51.50	46.8
WT into <i>mer^{kd}</i>	3.15	2.9
<i>mer^{kd}</i> into WT	0.98	1
Saline controls	1.15	–

Each experimental group included up to six mice per analysis. WT, wild type.

*The data is representative of two independent experiments.

downregulate the production of cytokines such as tumour-necrosis factor- α (TNF- α)¹⁸. In addition, growth arrest-specific gene 6 (GAS6), a ligand for the Axl/Tyro3/Mer family of receptor tyrosine kinases, has been shown to mediate the binding of phosphatidylserine to macrophages^{19,20}. These findings led us to speculate that Mer on macrophages may participate in the phagocytosis and clearance of apoptotic cells.

To examine whether Mer is involved in the clearance of apoptotic cells, we injected adult mice with dexamethasone to induce apoptosis of cortical thymocytes. Thymocytes were isolated and the percentage of apoptotic thymocytes was determined by flow cytometry of annexin-positive cells or by measuring DNA content. As shown in Table 1, the thymi of *mer^{kd}* mice showed a sevenfold increase in remnant apoptotic cells compared with the thymi of wild-type mice. Histological analysis of the thymus from wild-type mice treated with dexamethasone showed typical atrophy of the thymus, with the condensed nuclei of apoptotic cells mostly associated with macrophages (data not shown). In contrast, *mer^{kd}* mice treated with a similar dose of dexamethasone showed atrophy of the thymus but a greater number of apoptotic cells (32.38% compared with 4.4% in wild-type mice) with nuclear condensation

and pyknotic bodies was detected outside macrophages. Saline-treated wild-type and *mer^{kd}* mice showed negligible pyknotic bodies and apoptosis in the thymus. Although Mer is not expressed on thymocytes¹⁵, differences in the rate of dexamethasone-induced thymocyte death in the two genotypes was eliminated as an explanation for the increased presence of apoptotic cells in the dexamethasone-treated *mer^{kd}* mice. Thymocytes, isolated from both animals and treated with dexamethasone *in vivo*, were collected at hourly intervals and stained to measure the DNA content of the dying thymocyte population by flow cytometry²¹. Similar rates of cell death were observed in thymocytes from the *mer^{kd}* and wild-type animals (data not shown). Thus, thymocyte susceptibility to

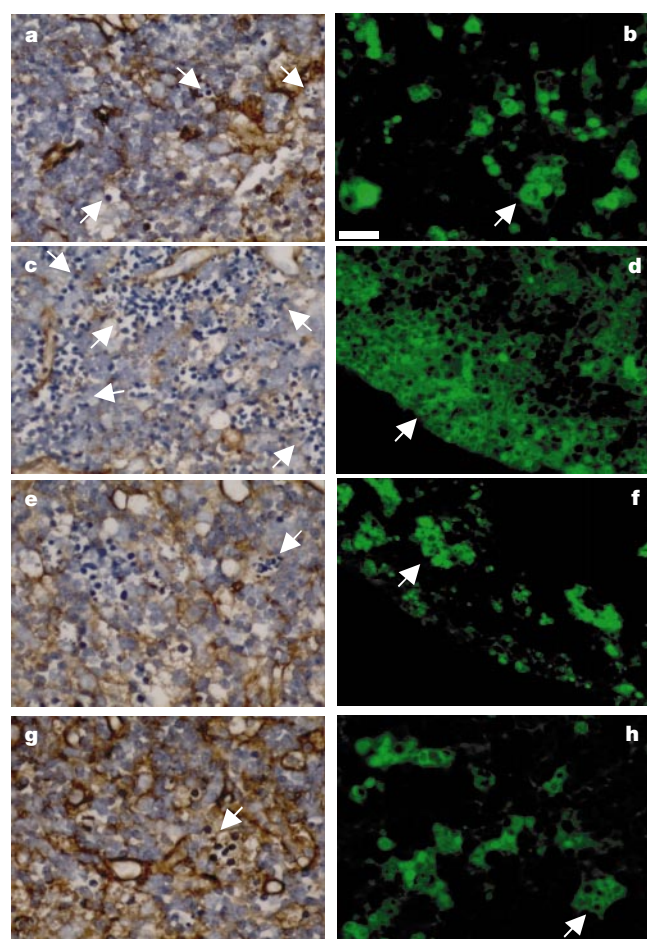


Figure 1 Effect of bone marrow reconstitution on dexamethasone treatment of thymus. **a, b**, Thymus from dexamethasone-treated wild-type into wild-type chimaeric (WT into WT) mouse stained for RCA-1 and apoptotic cells, respectively. **c, d**, Thymus from dexamethasone-treated *mer^{kd}* into *mer^{kd}* chimaeric mouse stained for RCA and apoptotic cells, respectively. **e, f**, Thymus from dexamethasone-treated WT into *mer^{kd}* chimaera stained for RCA-1 and apoptotic cells, respectively. **g, h**, Thymus from dexamethasone-treated *mer^{kd}* into WT chimaera stained for RCA-1 and apoptotic cells, respectively. Arrows, apoptotic and pyknotic bodies with nuclear condensation. Scale bar, 20 μ m. Observations were made using a Nikon fluorescent microscope.

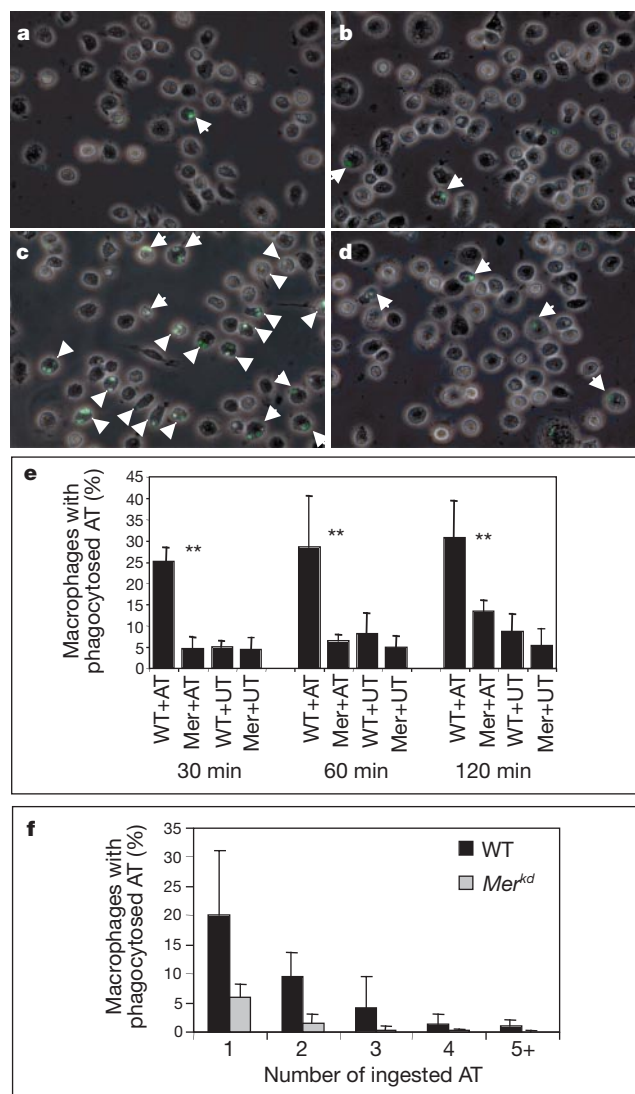


Figure 2 Macrophage phagocytosis of apoptotic thymocytes from *mer^{kd}* and wild-type mice. Apoptotic and untreated thymocytes were fluorescently labelled with CellTracker green to visualize phagocytosis (arrowheads). Fluorescence and phase-contrast images were superimposed. **a, b**, Phagocytosis of untreated thymocytes by macrophages from wild-type and *mer^{kd}* mice, respectively. **c, d**, Phagocytosis of apoptotic thymocytes by macrophages from wild-type and *mer^{kd}* mice, respectively. **e**, *In vitro* time course of phagocytosis comparing macrophages from wild-type and *mer^{kd}* mice. **f**, The percentage of macrophages that phagocytosed apoptotic thymocytes (AT) were plotted according to the number ingested. Four to five experiments were performed for each sample, and the mean and standard deviation is plotted. AT indicates macrophages that received apoptotic thymocytes; UT indicates samples that received untreated thymocytes. Double asterisk, $P < 0.001$.

apoptosis did not account for the differences observed in the dexamethasone-treated animals and indicated a deficiency in the clearance of apoptotic cells by *mer^{kd}* mice.

The *in vivo* deficiency in the clearance of apoptotic thymocytes was confirmed by two alternative approaches. First, we injected fluorescent-labelled apoptotic thymocytes into the peritoneal cavity after inducing macrophages with thioglycollate. One hour later, the extracted cells showed that more apoptotic thymocytes were cleared on average in wild-type mice than in *mer^{kd}* mice (6.9% versus 17.8% remaining apoptotic thymocytes, respectively). Similarly, using an intravenous injection of fluorescence-labelled syngeneic apoptotic lymphocytes and assessing the clearance of apoptotic cells within the spleen after 14 h by flow cytometry, we found that spleens from wild-type mice showed complete clearance of apoptotic cells; however, analysis of *mer^{kd}* mice revealed that approximately 2% of fluorescent-positive cells had failed to be phagocytosed. Hence, these results confirmed that *mer^{kd}* mice could not adequately phagocytose and clear cells *in vivo*.

To determine whether the clearance deficiency in the *mer^{kd}* mouse was due to macrophages, we created chimaeric mice by transferring bone marrow from donor mice into irradiated recipient mice. Histology and flow cytometric analysis were used to estimate the number of remnant apoptotic cells in the thymus of the chimaeric animals after dexamethasone treatment. Saline treatment of chimaeric wild-type mice reconstituted with syngeneic bone marrow from wild-type mice (wild type into wild type) showed rare apoptotic cells (Table 1). The wild-type into wild-type chimaeras treated with dexamethasone showed the expected clearance of apoptotic cells, with intense macrophage staining and compartmentalization of the apoptotic cells (Table 1, Fig. 1a, b). *Mer^{kb}* mice reconstituted with syngeneic bone marrow (*mer^{kd}* into *mer^{kd}*) had an abundance of apoptotic cells (46-fold increase compared with the wild-type mice) and negligible compartmentalization of the apoptotic thymocytes (Table 1, Fig. 1c, d). However, the *mer^{kd}* mice reconstituted with bone marrow from wild-type mice (wild type into *mer^{kd}*) showed clearance of apoptotic cells (3.15%) by macrophages and the compartmentalization of apoptotic cells (Table 1, Fig. 1e, f). Wild-type mice reconstituted with bone marrow from *mer^{kd}* (*mer^{kd}* into wild type) mice did not have normal phagocytosis (Table 1, Fig. 1g, h). This may be due to the presence of resident, irradiation-resistant, non-bone-marrow-derived phagocytes of the

wild-type mouse that compensated for the defective donor-derived *mer^{kd}* macrophages. These results confirmed that a bone-marrow-derived population, presumably normal macrophages, was responsible for correcting the clearance deficit in the *mer^{kd}* mice.

To determine whether primary macrophages from *mer^{kd}* mice were defective in clearance of apoptotic cells, peritoneal macrophages were collected and tested *in vitro* for their ability to ingest apoptotic thymocytes. As shown in Fig. 2, macrophages isolated from wild-type and *mer^{kd}* mice showed negligible phagocytosis of untreated thymocytes. However, macrophages from wild-type mice were competent in the phagocytosis of apoptotic thymocytes over a 2-h time course as previously reported (Fig. 2c, e)⁶. In contrast, macrophages from *mer^{kd}* mice were markedly deficient in phagocytosis of apoptotic thymocytes, particularly at 30 and 60 min (Fig. 2d, e). Macrophages from *mer^{kd}* mice were 83–94% less efficient in phagocytosis of apoptotic cells than macrophages from wild-type mice. Furthermore, the macrophages that ingested apoptotic thymocytes had an average of one to three apoptotic thymocytes with as many as ten particles in macrophages from wild-type mice (Fig. 2f). In contrast, macrophages from *mer^{kd}* mice had a negligible number of phagocytosed apoptotic thymocytes (Fig. 2f).

Binding assays were performed to determine whether the deficiency in phagocytosis of apoptotic thymocytes was due to an inability of macrophages to bind to apoptotic cells. As shown in Fig. 3a, b macrophages from both wild-type and *mer^{kd}* mice bound to apoptotic thymocytes similarly with $25.66 \pm 0.75\%$ and $25.87 \pm 1.83\%$, respectively. Using scanning electron microscopy (SEM), the average number of bound apoptotic thymocytes per macrophage was determined by random selection of the macrophage population. After 60 min of incubation, wild-type macrophages bound 1.78 ± 0.24 apoptotic thymocytes per macrophage ($n = 45$), and similar binding was observed for macrophages from *mer^{kd}* mice with an average of 2.48 ± 0.36 apoptotic thymocytes bound per macrophage ($n = 45$). Equivalent binding was also confirmed after 30 min of incubation. These averages were lower than those previously reported and may reflect differences in our experimental protocols²². Consistent with equivalent binding of apoptotic thymocytes, there was no significant difference in the CD14 protein levels in western blot analysis of macrophages from wild-type and *mer^{kd}* mice (data not shown). Under phagocytic conditions, SEM revealed that macrophages from wild-type mice proceeded to ingest surface apoptotic thymocytes after 60 min incubation, at which point only a few apoptotic thymocytes were bound to the surface of macrophages (Fig. 3c). In contrast, under these conditions, macrophages from *mer^{kd}* mice had only bound

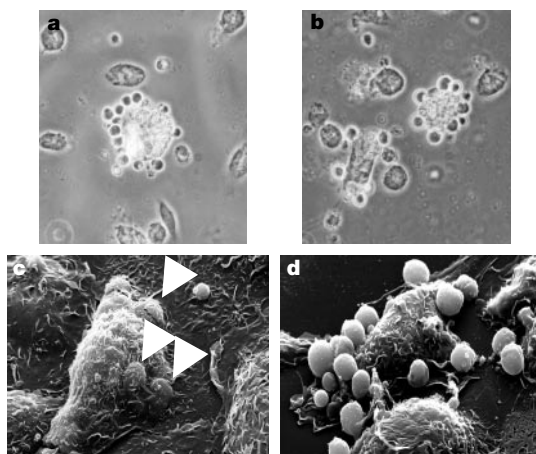


Figure 3 Binding of apoptotic thymocytes to wild-type and *mer^{kd}* macrophages. **a, b**, Wild-type and *mer^{kd}* macrophages with bound apoptotic thymocytes after 60 min of incubation, respectively ($\times 60$ magnification). **c**, SEM micrograph analysis of macrophage from wild-type mouse that phagocytosed apoptotic thymocytes (arrowheads; scale bar, 10 μ M). **d**, SEM micrograph of macrophages from *mer^{kd}* mice with bound apoptotic thymocytes that could not phagocytose these apoptotic cells. Scale bar, 10 μ M.

Table 2 Phagocytosis of *L. monocytogenes* and latex beads by wild-type and *mer^{kd}* macrophages

	Wild-type macrophages with ingested particles (%)	<i>mer^{kd}</i> macrophages with ingested particles (%)
<i>L. monocytogenes</i>		
Experiment 1	55.1	62.2
Experiment 2	92.8	99.6
2 μ m carboxylate beads	80.77 ± 12.24	82.03 ± 12.41
+50 μ M cytochalasin B	17.55 ± 6.01	12.41 ± 8.70
6 μ m carboxylate beads	18.00 ± 7.43	22.14 ± 7.21
+50 μ M cytochalasin B	04.72 ± 2.32	10.90 ± 3.97
6 μ m BSA-conjugated carboxylate beads	38.25 ± 18.75	50.47 ± 15.19
+50 μ M cytochalasin B	18.11 ± 11.47	16.11 ± 7.33
6 μ m anti-BSA/BSA-conjugated carboxylate beads	34.17 ± 16.19	24.79 ± 13.93
+50 μ M cytochalasin B	10.51 ± 3.08	7.20 ± 2.13
6 μ m IgG-conjugated carboxylate beads	44.34 ± 1.55	51.64 ± 11.96
+50 μ M cytochalasin B	31.00 ± 1.66	39.80 ± 10.00

Phagocytosis of latex beads represents an average of up to four independent experiments.

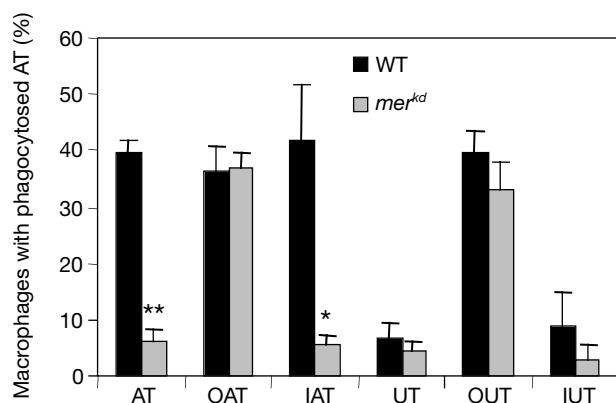


Figure 4 Phagocytosis of opsonized thymocytes. Dexamethasone-induced apoptotic (AT) and untreated thymocytes (UT) were opsonized with anti-CD3 or isotype-matched antibody. OAT, AT opsonized with anti-CD3; IAT, AT incubated with isotype control antibody; OUT, UT with anti-CD3; IUT, UT with isotype-matched antibody. The average and standard error from three experiments are plotted. Black bars, phagocytosis by macrophages from wild-type mice; light grey bars, phagocytosis by macrophages from *mer^{kd}* mice. Asterisk, $P < 0.05$; double asterisk, $P < 0.001$.

apoptotic thymocytes, and negligible phagocytosis was visible (Fig. 3d). Thus, it seems that the deficiency in phagocytosis of apoptotic thymocytes cannot be attributed to deficits in macrophage recognition or binding. Rather, these data indicate that the cytoplasmic domain of the Mer receptor may be responsible for triggering phagocytosis of apoptotic cells.

To determine whether *mer^{kd}* macrophages had a general defect in phagocytosis, we tested macrophages from *mer^{kd}* mice for their ability to ingest bacteria, latex beads and opsonized particles including opsonized latex beads. There were no differences in the ingestion of *Listeria monocytogenes* or carboxylate-coated latex beads (2 μ m and 6 μ m) between macrophages from *mer^{kd}* and wild-type mice (Table 2). As the 6- μ m latex beads are about the same size as the apoptotic thymocytes, these results indicate that the deficit in phagocytosis of apoptotic thymocytes was not due to the size of the thymocytes. To further address other deficits in phagocytosis, Fc-receptor-mediated phagocytosis was tested using 6- μ m carboxylate-coated beads either directly conjugated to immunoglobulin G or to bovine serum albumin (BSA) with a subsequent binding of anti-BSA IgG antibody. As shown in Table 2, there was no significant deficiency in the phagocytosis of the opsonized latex particles by macrophages isolated from *mer^{kd}* mice compared with those from wild-type mice. Treatment with cytochalasin B, an inhibitor of phagocytosis that allows discrimination of nonspecific interactions with macrophages, resulted in a marked decrease in the phagocytosis of *L. monocytogenes* (data not shown) and carboxylate-coated latex beads in both the wild-type and *mer^{kd}* macrophages (Table 2). These data show that macrophages from *mer^{kd}* mice were fully competent in the phagocytosis of particles other than apoptotic cells.

To demonstrate that the Mer-mediated phagocytosis of apoptotic cells by macrophages was independent of Fc-receptor-mediated uptake, we opsonized apoptotic thymocytes with anti-

CD3 antibodies and examined the phagocytosis of these particles. Macrophages from *mer^{kd}* mice were able to ingest the opsonized apoptotic thymocytes and opsonized untreated thymocytes similarly to macrophages from wild-type mice (Fig. 4). However, the *mer^{kd}* phagocytic deficiency was still evident in the control experiments with non-opsonized apoptotic thymocytes and in apoptotic thymocytes treated with an isotype control antibody (Fig. 4). These results showed that the phagocytic deficiency in the macrophages from *mer^{kd}* mice was specifically restricted to the engulfment of apoptotic thymocytes, and was not due to a size-dependent phenomenon or a general deficiency in phagocytosis.

A major consequence of the deficiency to clear apoptotic cells may be autoimmunity^{3,4,12}. Systemic exposure to apoptotic thymocytes in mice results in the transient production of autoantibodies, and defects in clearance and phagocytosis have been correlated to increased autoantibodies^{23,24}. Given the deficiency in clearing apoptotic cells, it was intriguing to find that greater than 50% of four- to five-month-old *mer^{kd}* mice had a higher level of autoantibodies (similar to *MLR/lpr* mice) against nuclear components than wild-type mice (Table 3). We suggest that the *mer^{kd}* mice and the inability to adequately clear apoptotic cells may be a suitable model for systemic lupus erythematosus. Moreover, a slight increase in the amount of small pyknotic bodies was detected in the spleens of older *mer^{kd}* mice compared with wild-type mice. However, pyknotic bodies are rare in the unchallenged *mer^{kd}* thymus, indicating that alternative cell types, such as thymic nurse cells^{23–25} and other mechanisms may be able to compensate for Mer in the removal of apoptotic thymocytes. Recently, a rat model with retinal degeneration where failure of retinal pigment epithelial cells to phagocytose the shed outer segment layers results in the death of photoreceptor cells, and human patients with retinitis pigmentosa have been shown to contain mutations in the *mer* gene^{26,27}. The functional mutation in Mer reported here also renders the *mer^{kd}* mice blind and this implies that other phagocytes and apoptotic tissues are affected by Mer. Moreover, the *Axl/Mer/Tyro3* triple knockout mice had several abnormalities that included blindness, neurological abnormalities and splenomegaly owing to increased numbers of apoptotic cells, this increase could result from a lack of clearance of apoptotic cells²⁸. Thus, the removal of apoptotic cells mediated partly by the cytoplasmic signalling domain of Mer may be critical to the maintenance of tissue homeostasis and the prevention of autoimmunity. □

Methods

In vivo clearance of apoptotic cells, immunohistochemistry and flow cytometry

The *mer^{kd}* mice were crossed for six generations to C57BL/6 background, and C57BL/6 (Jackson Labs) were used as wild-type controls. Adult mice were injected intraperitoneally with 0.2 mg dexamethasone sodium phosphate (American Reagent Laboratories) per 25 g. Mice injected with 1 $\times$ PBS were used as controls. After 24 h, mice were killed and a lobe from each thymus was fixed in 10% buffered formalin (Polysciences) and embedded in paraffin. Sections (5 μ m) were stained for RCA-1, a lectin present on macrophages, or apoptosis. RCA-1 staining was modified from the described protocol²⁹. Briefly, deparaffinized sections were treated with 2% hydrogen peroxide in methanol for 5 min at 43 °C. All incubations were performed at 43 °C. Subsequently, sections were treated with 0.025% protease (Sigma type XXVII) for 2 min, rinsed in 1 $\times$ PBS and blocked with 1 $\times$ PBS containing 0.1% Triton X-100 and 0.1% BSA for 5 min. Biotinylated RCA-1 (Sigma) was used at a dilution of 1:500 in blocking solution, and incubated for 30 min. Streptavidin-peroxidase complex was made using the ABC kit (Vector), and peroxidase activity was detected using diaminobenzidine (DAB, Vector). Staining for apoptotic cells was performed using the NeuroTACS kit (Trevigen) with FITC/streptavidin to detect the biotinylated nucleotides incorporated into apoptotic cells by Klenow enzyme. The second lobe of each thymus was mechanically dissociated. Unfixed cells were stained with Annexin/FITC (Pharmingen). Ethanol-fixed cells were stained with propidium iodide as described²¹. The percentage of cell death was determined on a Becton Dickinson FACSscan using Cytomation software. Two independent experiments were analysed with up to six mice per group. The standard error was within 20% of the mean. A second *in vivo* clearance assay used mice that were injected with 3 ml of thioglycollate to induce macrophages. After 3 days, about 4 $\times$ 10⁶ dexamethasone-induced, CellTracker green-labelled, apoptotic thymocytes were injected and cells were collected after 1 h. Cells were fixed and the remaining apoptotic thymocytes were quantified by flow cytometry as a

Table 3 Elevated anti-DNA antibodies in sera from *mer^{kd}* mice compared with wild-type mice

Genotype	Number of mice with positive anti-DNA sera
Wild-type controls	0/16
<i>mer^{kd}</i>	14/24*
<i>MLR/lpr</i>	2/2

* Student's *t*-test comparing wild-type and *mer^{kd}* mice ($P < 0.001$).

percentage of the total number of peritoneal cells counted (duplicated experiment, $n = 6$ per genotype). In the third *in vivo* clearance assay a 1:1 mixture of splenocytes and thymocytes from *mer^{kd}* and wild-type mice was used stained with 50 μg of 5-carboxy-tetramethylrhodamine, succinimidyl ester (TAMRA, SE; Molecular Probes) for 30 min at 37 °C and γ -irradiated (1,500 rad) to induce apoptosis. Four hours after irradiation, 8×10^7 apoptotic cells per mouse were injected intravenously. Mice were killed 14 h after injection, and the isolated spleen cells were analysed by flow cytometry for TAMRA labelling.

Bone-marrow reconstitution

Mice 4- to 5-weeks old were irradiated with 800 centigrays (cGy) at a dose rate of 50 cGy min^{-1} using an AECL theratron 60 Cobalt teletherapy machine. Mice were reconstituted by intravenous injection with $2-4 \times 10^7$ bone marrow cells from donor mice in 500 μl 1 $\times$ PBS. The chimaeric mice were allowed to reconstitute for more than 4 months before being challenged with dexamethasone. Bone marrow from mice expressing the green fluorescent protein (GFP) under the MHC class I promoter (provided by R. Tisch and J. Frelinger) was injected to determine the amount of reconstitution in the mice. Flow cytometry of blood leukocytes showed around 90% reconstitution.

Phagocytosis assay

Peritoneal exudate cells (PECs) were collected from mice 3 days after injection with aged 3% thioglycollate (Difco). Around 2×10^5 PECs were plated onto 12-well glass coverslips in RPMI medium (Gibco) supplemented with 5% fetal bovine serum (FBS Irvine Scientific), 50 units penicillin G (Gibco) and 50 μg ml^{-1} streptomycin sulphate (Gibco). The phagocytosis procedure was modified from previously described methods⁵. Briefly, cells were allowed to attach for a minimum of 2 h, washed with 1 $\times$ PBS and left in medium overnight. The next day, thymocytes were harvested from 3- to 4-week-old C57BL/6 mice. Around 5×10^6 thymocytes per ml were treated with 2 μM dexamethasone (Sigma) for incubation period. Untreated thymocytes were incubated for the same period without dexamethasone treatment. Some cell death was observed in the untreated population after the incubation period. After dexamethasone treatment, cells were fluorescein-labelled with CellTracker green (Molecular Probes). Cells were incubated at 37 °C for 30 min with 2 μM CellTracker green resuspended in complete medium, washed with 1 $\times$ PBS, and incubated in medium for an additional 30 min. Cells were washed again with 1 $\times$ PBS, and resuspended in complete medium. This treatment routinely yielded over 70% apoptotic thymocytes as measured by propidium iodide staining of the cells. Apoptosis of the thymocytes was also determined by DNA laddering and annexin/propidium iodide staining. Fluorescent thymocytes in 20-fold excess (2×10^6 cells) were plated onto macrophages for various times. To opsonize apoptotic thymocytes, they were treated with dexamethasone for 4 h, followed by opsonization with anti-CD3 or isotype-matched antibody (rat IgG2b, Pharmingen, 1 μg per 10^6 cells), and labelled with CellTracker as described above. Opsonization was confirmed by flow cytometry. *Listeria monocytogenes* was fluorescein-labelled with CellTracker green as described above, except that labelling was performed in Luria broth media (Difco). Carboxylate-coated beads (2 μm from Molecular Probes, and 6 μm from Polysciences) were washed three times with 1 $\times$ PBS and an equivalent of 2–5 μl of 2% bead mixture was plated onto PECs. Conjugation of the carboxylate-coated latex beads was performed as directed by Molecular Probes using EDAC reagent to crosslink the proteins to the beads. Opsonization of the beads was verified by ELISA detection. The wells were washed (five times) vigorously with 1 $\times$ PBS to remove most non-phagocytosed material. Cells were fixed with 1% paraformaldehyde (Sigma), and stored in the dark at 4 °C before microscopic observation. Macrophages were counted at random, and phagocytosis was visualized and quantified using a Nikon fluorescent microscope and Scion image analysis software. Each experiment was duplicated. Data were analysed (Fig. 2e) with a factorial analysis of variance using genotype, treatment type of thymocytes and time as factors. Results indicated a small but significant effect of time ($F(2, 36) = 3.484$; $P = 0.041$). There was a significant interaction between genotype and treatment ($F(1, 36) = 34.88$; $P < 0.001$). Multiple comparisons were conducted with Tukey's test to interpret the two-way interaction. The wild-type macrophages that received apoptotic thymocytes were significantly different from each of the other three groups ($P < 0.001$) at each time point. Each of the other three groups did not differ significantly from each other. Student's *t*-test was also applied as a statistical determinant.

Binding of thymocytes

Thymocyte binding to macrophages was performed as described in the phagocytosis procedure except for the following changes: (1) untreated and apoptotic thymocytes were not fluorescein-labelled as binding could be visually scored; (2) binding was conducted under low serum (0.2%)²² and in the absence of serum (binding was observed under both conditions); (3) cells were washed gently at least eight times with 1 ml PBS; and (4) untreated thymocytes were used immediately after isolation to reduce background due to dead cells detected after 6 h of incubation. Macrophages were randomly counted and the quantification of binding of apoptotic thymocytes to macrophages was reported as the mean and standard error.

Scanning electron microscopy

Cells grown on glass coverslips were fixed with 3% glutaraldehyde in 0.1 M sodium cacodylate buffer and stored at 4 °C until processed. Cells were washed for 20 min each with 1% osmium tetroxide, then by 1% tannic acid, and 1% osmium tetroxide. After

dehydration, coverslips were coated with 20 nm of Au:Pd (60:40) and viewed on a Cambridge S200 SEM (LEO Electron Microscopy).

Autoantibody detection

Antibodies to nuclear antigens (ANA) were determined by the Crithidia luciliae assay³⁰ crithidia slides from Kallestad. Serum from 4- to 5-month-old mice was tested blind, and the level of immunofluorescence was scored from 0 to 4 using an Olympus microscope.

Received 5 January; accepted 14 March 2001.

- Jacobson, M. D., Weil, M. & Raff, M. C. Programmed cell death in animal development. *Cell* **88**, 347–354 (1997).
- Savill, J., Fadok, V., Henson, P. & Haslett, C. Phagocyte recognition of cells undergoing apoptosis. *Immunol. Today* **14**, 131–136 (1993).
- Casiano, C. A. & Tan, E. M. Recent developments in the understanding of antinuclear autoantibodies. *Int. Arch. Allergy Immunol.* **111**, 308–313 (1996).
- Rosen, A. & Casciola-Rosen, L. Autoantigens as substrates for apoptotic proteases: implications for the pathogenesis of systemic autoimmune disease. *Cell Death Diff.* **6**, 6–12 (1999).
- Platt, N., Suzuki, H., Kurihara, Y., Kodama, T. & Gordon, S. Role for the class A macrophage scavenger receptor in the phagocytosis of thymocytes *in vitro*. *Proc. Natl Acad. Sci. USA* **93**, 12456–12460 (1996).
- Fadok, V. A. *et al.* Exposure of phosphatidylserine on the surface of apoptotic lymphocytes triggers specific recognition and removal by macrophages. *J. Immunol.* **148**, 2207–2216 (1992).
- Fadok, V. A. *et al.* A receptor for phosphatidylserine-specific clearance of apoptotic cells. *Nature* **405**, 85–90 (2000).
- Savill, J., Dransfield, I., Hogg, N. & Haslett, C. Vitronectin receptor-mediated phagocytosis of cells undergoing apoptosis. *Nature* **343**, 170–173 (1990).
- Savill, J., Hogg, N., Ren, Y. & Haslett, C. Thrombospondin cooperates with CD36 and the vitronectin receptor in macrophage recognition of neutrophils undergoing apoptosis. *J. Clin. Invest.* **90**, 1513–1521 (1992).
- Stern, M., Savill, J. & Haslett, C. Human monocyte-derived macrophage phagocytosis of senescent eosinophils undergoing apoptosis. Mediation by α v β 3/CD36/thrombospondin recognition mechanism and lack of phagocytic response. *Am. J. Pathol.* **149**, 911–921 (1996).
- Devitt, A. *et al.* Human CD14 mediates recognition and phagocytosis of apoptotic cells. *nature* **392**, 505–509 (1998).
- Taylor, P. R. *et al.* A hierarchical role for classical pathway complement proteins in the clearance of apoptotic cells *in vivo*. *J. Exp. Med.* **192**, 359–366 (2000).
- O'Bryan, J. P. *et al.* axl, a transforming gene isolated from primary human myeloid leukemia cells, encodes a novel receptor tyrosine kinase. *Mol. Cell Biol.* **11**, 5016–5031 (1991).
- Lai, C., Gore, M. & Lemke, G. Structure, expression, and activity of Tyro 3, a neural adhesion-related receptor tyrosine kinase. *Oncogene* **9**, 2567–2578 (1994).
- Graham, D. K. *et al.* Cloning and development expression analysis of the murine c-mer tyrosine kinase. *Oncogene* **10**, 2349–2359 (1995).
- McCloskey, P. *et al.* GAS6 mediates adhesion of cells expressing the receptor tyrosine kinase Axl. *J. Biol. Chem.* **272**, 23285–23291 (1997).
- Fridell, Y. W., Villa, J. J., Attar, E. C. & Liu, E. T. GAS6 induces Axl-mediated chemotaxis of vascular smooth muscle cells. *J. Biol. Chem.* **273**, 7123–7126 (1998).
- Camenisch, T. D., Koller, B. H., Earp, H. S. & Matsushima, G. K. A novel receptor tyrosine kinase, Mer, inhibits TNF- α production and lipopolysaccharide-induced endotoxic shock. *J. Immunol.* **162**, 3498–3503 (1999).
- Nagata, K. *et al.* Identification of the product of growth arrest-specific gene 6 as a common ligand for Axl, Sky, and Mer receptor tyrosine kinases. *J. Biol. Chem.* **271**, 30022–30027 (1996).
- Nakano, T. *et al.* Cell adhesion to phosphatidylserine mediated by a product of growth arrest-specific gene 6. *J. Biol. Chem.* **272**, 29411–29414 (1997).
- Reap, E. A. *et al.* Radiation and stress-induced apoptosis: A role for Fas/Fas ligand interactions. *Proc. Natl Acad. Sci. USA* **94**, 5750–5755 (1997).
- Licht, R., Jacobs, C. W. M., Tax, W. J. M. & Berden, J. H. M. An assay for the quantitative measurement of *in vitro* phagocytosis of early apoptotic thymocytes by murine resident peritoneal macrophages. *J. Immunol. Methods* **223**, 237–248 (1999).
- Mervorach, D., Zhou, J. L., Song, X. & Elkon, K. B. Systemic exposure to irradiated apoptotic cells induces autoantibody production. *J. Exp. Med.* **188**, 387–392 (1998).
- Meryhew, H. L., Shaver, C., Messner, R. P. & Runquist, O. A. Mononuclear phagocyte system dysfunction in murine SLE: abnormal clearance kinetics precede clinical disease. *J. Lab. Clin. Med.* **117**, 181–193 (1991).
- Aguilar, L. K., Aguilar-Cordova, E., Cartwright, J. & Belmont, J. W. Thymic nurse cells are sites of thymocyte apoptosis. *J. Immunol.* **152**, 2645–2651 (1994).
- D'Cruz, P. M. *et al.* Mutation of the receptor tyrosine kinase gene MerTK in the retinal dystrophic RCS rat. *Human Mol. Genet.* **9**, 645–651 (2000).
- Gal, A. *et al.* Mutations in MERTK, the human orthologue of the RCS rat retinal dystrophy gene, cause retinitis pigmentosa. *Nature Med.* **26**, 270–271 (2000).
- Lu, Q. *et al.* Tyro-3 family receptors are essential regulators of mammalian spermatogenesis. *Nature* **398**, 723–728 (1999).
- Hsu, S. & Ree, H. J. Histochemical studies on lectin binding in reactive lymphoid tissues. *J. Histochem. Cytochem.* **31**, 538–546 (1983).
- Eisenberg, R. A., Winfield, J. B. & Cohen, P. L. Subclass restriction of anti-Sm antibodies in MRL mice. *J. Immunol.* **129**, 2146–2149 (1982).

Acknowledgements

We thank R. Bagnell and V. Madden for assistance with microscopy, C. Lawler for statistical analysis, E. Cash and T. Lee for histological assistance, S. Parker for irradiating mice and L. Su for critical reading of the manuscript.

Correspondence and requests for materials should be addressed to G.K.M. (e-mail: gkmat@med.unc.edu).

Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*

Asuka Miura^{*†}, Shoji Yonebayashi^{†‡}, Koichi Watanabe[§], Tomoko Toyama[§], Hiroaki Shimada^{||} & Tetsuji Kakutani^{*‡§}

^{*} National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

[‡] National Institute of Agrobiological Resources, Tsukuba, Ibaraki 305-8602, Japan

[§] CREST, Japan Science and Technology Corporation, Tokyo 101-0062, Japan

^{||} Science University of Tokyo, Noda, Chiba 278-8510, Japan

[†] These authors contributed equally to this work.

A major component of the large genomes of higher plants and vertebrates comprises transposable elements and their derivatives, which potentially reduce the stability of the genome¹. It has been proposed that methylation of cytosine residues may

suppress transposition, but experimental evidence for this has been limited^{2–5}. Reduced methylation of repeat sequences results from mutations in the *Arabidopsis* gene *DDM1* (*decrease in DNA methylation*)⁶, which encodes a protein similar to the chromatin-remodelling factor SWI2/SNF2 (ref. 7). In the *ddm1*-induced hypomethylation background, silent repeat sequences are often reactivated transcriptionally, but no transposition of endogenous elements has been observed^{8–11}. A striking feature of the *ddm1* mutation is that it induces developmental abnormalities by causing heritable changes in other loci^{12,13}. Here we report that one of the *ddm1*-induced abnormalities is caused by insertion of *CAC1*, an endogenous CACTA family transposon. This class of *Arabidopsis* elements transposes and increases in copy number at high frequencies specifically in the *ddm1* hypomethylation background. Thus the *DDM1* gene not only epigenetically ensures proper gene expression^{13–16}, but also stabilizes transposon behaviour, possibly through chromatin remodelling or DNA methylation.

To study the molecular basis for *ddm1*-induced developmental abnormalities, we identified the mutated gene responsible for one of them, *clam* (*clm*), which is characterized by lack of elongation in shoots and petioles¹² (Fig. 1a). This phenotype was initially unstable, and phenotypically normal sectors were occasionally observed (Fig. 1b–f). In subsequent generations, the *clm* phenotype stabilized in some of the progeny families and was inherited as a recessive mendelian trait, which could be mapped genetically¹². By genotyping 926 chromosomes from a mapping cross, we narrowed the *clm* locus to a 64-kilobase (kb) region in bacterial artificial chromosome (BAC) clone T3A5 (GenBank AL132979) on chromosome 3. This region contains the *DWF4* gene, encoding 22- α -hydroxylase, a 513-amino-acid protein that mediates the biosynthesis of brassinosteroid, a regulator of cell elongation¹⁷. Complementation tests indicated that *clm* is allelic to the *dwf4-4* mutation (see Supplementary Information). The sequence of the *DWF4* gene in the stable *clm* plant revealed a 4-base-pair (bp) insertion mutation in the second exon at nucleotide +527 from the translation start site. This insertion converts TAG to TAGC-TAG, creating a stop codon after the 149th amino acid. Although this defect is probably the cause of the stable *clm* phenotype, it cannot account for the instability of the *clm* phenotype in initial generations.

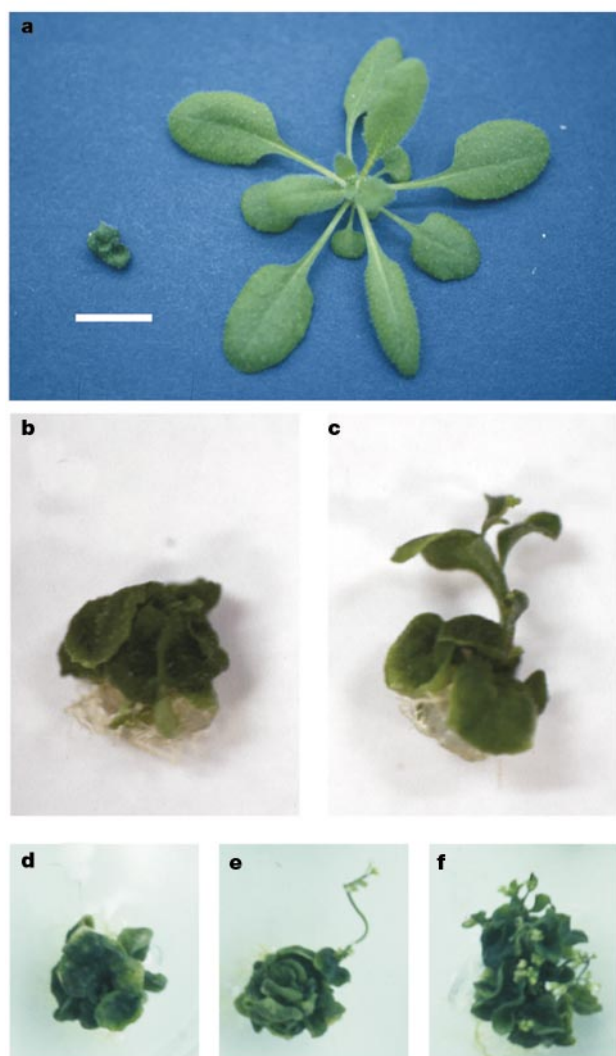


Figure 1 Stable and unstable *clm* phenotypes. **a**, A wild-type Columbia plant showing expanded leaves (right) and a *clm* homozygous plant with a defect in leaf and stem elongation (left). Both are 35 days old. **b–f**, *clm* homozygous plants with (**c**, **e**, **f**) or without (**b**, **d**) reversion sectors. Plants were grown on soil (**a**) or nutrient agar plates¹² (**b–f**). Scale bar, 10 mm.

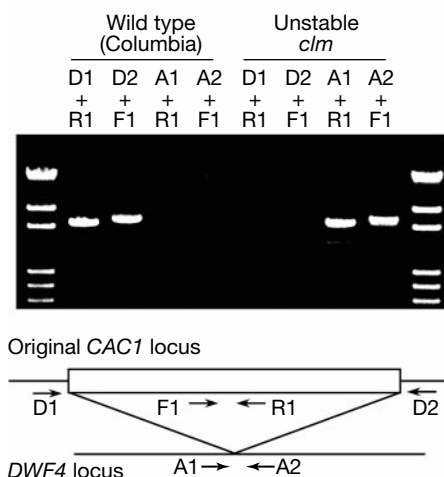


Figure 2 Transposition of *CAC1* element from the original donor site to the *DWF4* locus. Arrows indicate primers used for PCR. The DNA length markers are 19.3, 7.74, 5.53, 3.14, 2.69 and 2.32 kb. The sizes of transposon and the starting position (but not the length) of arrows for primers reflect the physical length.

In the unstable *clm* plant, we used genomic Southern analysis to identify a several-kilobase insertion in the *DWF4* gene (data not shown). We amplified the inserted DNA from unstable *clm* plants using the suppression polymerase chain reaction (PCR) technique¹⁸. Sequences obtained from each side of the insertion (523 and 154 nucleotides) showed exact matches to the 3' and 5' ends of a 8,479-bp sequence in chromosome 2 (nucleotides 52,296 to 60,774 in GenBank AC005897). No other exact match was found in the *Arabidopsis* genome database (<http://www.arabidopsis.org/blast/>). This 8,479 sequence has features typical of the CACTA family of transposons, which include maize *Spm/En* and snapdragon *Tam1* elements: the 5' and 3' ends have the conserved terminal inverted sequence CACTACAA, and the internal region includes a predicted gene (GenBank AAC97237) with similarity to the putative snapdragon transposase Tnp2 (ref. 19). To confirm the transposition of this sequence, we amplified two DNA fragments covering the entire element using internal sequences and the *DWF4* gene sequence as primers (Fig. 2). The estimated length of the product and the restriction digestion pattern were identical for the original copy and the insert in the *DWF4* gene, suggesting that the full-length element had transposed from chromosome 2 to the *DWF4* gene on chromosome 3. This transposon, designated CACTA1 (*CAC1*), was responsible for the unstable *clm* phenotype;

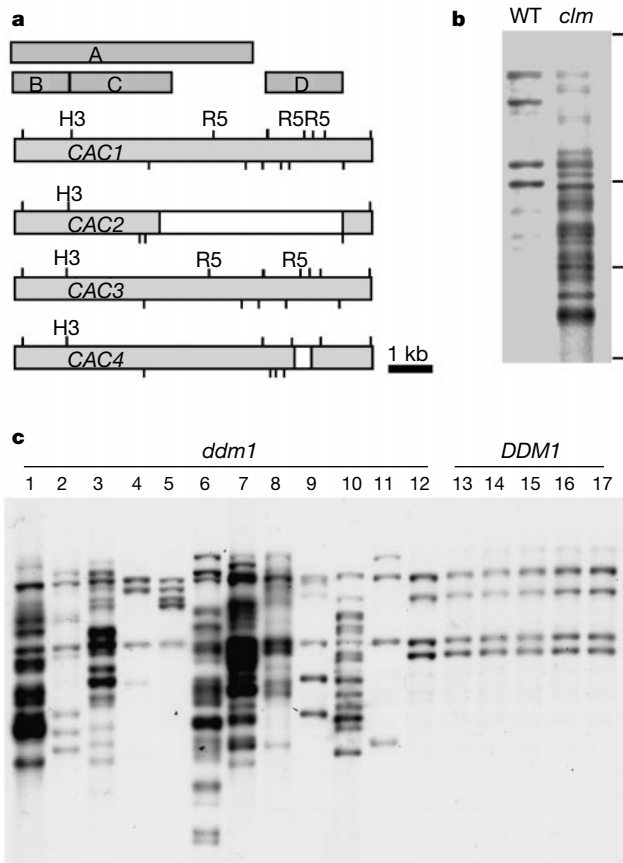


Figure 3 Transposition of *CAC* elements in the *ddm1* mutant background. **a**, The *CAC* transposon family. Open boxes in *CAC2* and *CAC4* show the internal deletions. Vertical bars show the restriction sites: H3, *HindIII*; R5, *EcoRV*; other bars above the elements, *HhaI*; below the elements, *HpaII*. The transpositions of *CAC* elements were examined by Southern blot analysis using *EcoRV* and probe A. **b**, *clm* and wild-type Columbia (WT) plants. **c**, The *ddm1* and wild-type *DDM1* lines self-pollinated six to seven times in parallel¹². Five out of 12 *DDM1* lines are shown. The DNA length markers are 19.3, 7.74, 5.53 and 4.25 kb.

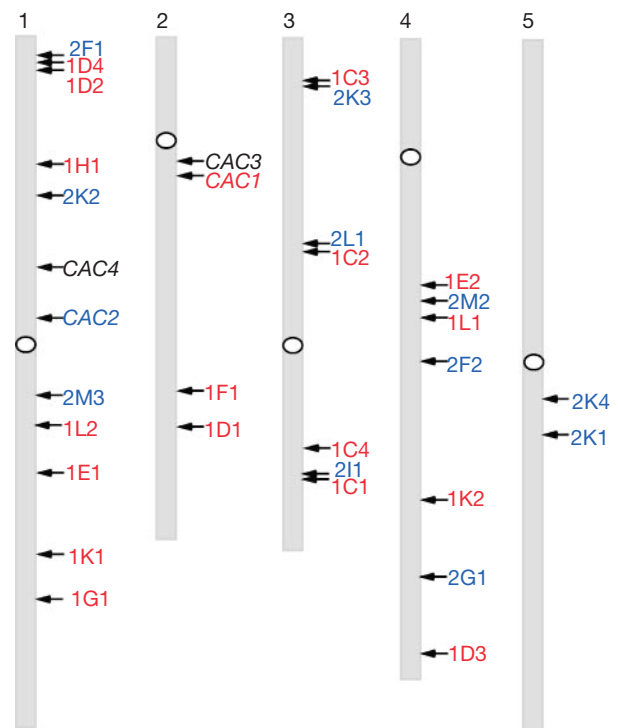


Figure 4 Transposition of *CAC1* and *CAC2* elements to unlinked loci. Sequences flanking the transposed copies were determined by suppression PCR (for determined sequences and methods see Supplementary Information). The red and blue loci show *CAC1* and *CAC2* copies, respectively. The transposed copies with second character D, E, F, G, H, I, K, L, M and C are from the *ddm1* lines shown in lanes 10, 9, 8, 7, 6, 5, 3, 2 and 1 of Fig. 3c and the *clm* line (Fig. 3b), respectively.

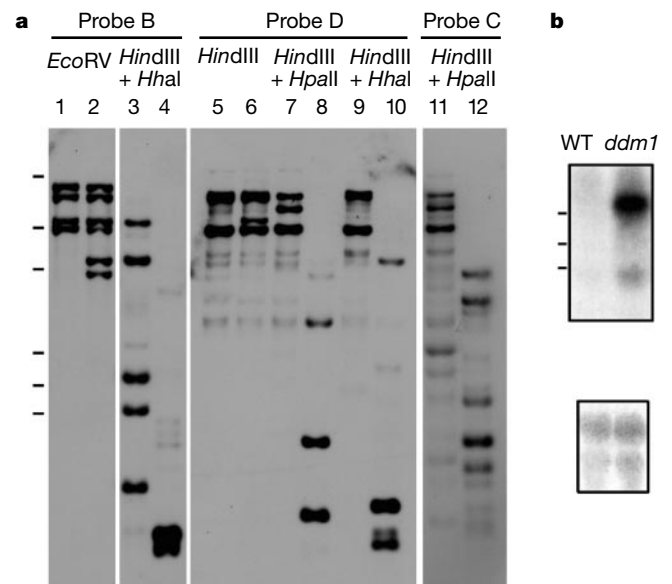


Figure 5 The *ddm1*-specific hypomethylation and RNA accumulation of *CAC* elements. **a**, Methylation-sensitive restriction enzymes (*HpaII* or *HhaI*) and probes B, C, D (Fig. 3a) were used to compare the methylation status of *CAC* elements between *ddm1* (even lanes) and Columbia wild-type (odd lanes) plants. The *ddm1* plant is before the repeated self-pollination (four generations before the plant shown in lane 10 of Fig. 3c). It still keeps the donor copies of *CAC* elements (lane 2). The DNA length markers are 19.3, 7.74, 5.53, 3.14, 2.69 and 2.32 kb. **b**, RNA blot analysis. Probe A (Fig. 3a) was used to detect *CAC* transcript in wild-type and *ddm1* (*clm*) plants. The RNA length markers are 6, 4 and 3 kb. Bottom panel, ribosomal RNA on the filter stained with methylene blue.

in the sector without the *clm* phenotype, restoration of the insertion site to the normal structure was found by sequencing PCR-amplified templates (see Supplementary Information).

The *Arabidopsis* genome contains three additional sequences distinct from but significantly similar to *CAC1*; their terminal 200 bp are 90–98% identical (Fig. 3a). As they are likely to comprise a new transposon family with *CAC1*, we designated them *CAC2* (GenBank AC069160, nucleotides 34,626–38,790), *CAC3* (GenBank AC006429, nucleotides 85,404–76,949) and *CAC4* (GenBank AC027135, nucleotides 72,903–64,858). Genomic Southern analysis using the methylation-insensitive restriction enzyme *EcoRV* revealed that *CAC* elements transposed and increased in the copy number in the *clm* line (Fig. 3b).

To see whether the mobilization of the *CAC* elements is a general phenomenon in the *ddm1* background, we examined genomic DNA of twelve *ddm1* lines that had independently and randomly self-pollinated six to seven times¹² (Fig. 3c). Eleven out of twelve *ddm1* lines exhibited changes in the band pattern of *CAC* elements, and six showed several-fold increases in band number (up to >20). Transposition of *CAC1* and *CAC2* was confirmed by sequencing the genomic regions flanking them. Both were found to transpose to unlinked loci throughout the genome (Fig. 4). In contrast, we never observed the change in *EcoRV* banding pattern for the *CAC* elements in any of the 12 control wild-type *DDM1* lines that had independently self-pollinated seven times in parallel (Fig. 3c), indicating that the mobilization of *CAC* elements is a consequence of the *ddm1* mutation.

Mobilization of transposons is often associated with hypomethylation and transcriptional activation^{20–23}. Indeed, the *CAC* elements were hypomethylated and transcriptionally activated in *ddm1* plants (Fig. 5), like other repeated sequences^{8–11}. PCR with reverse transcriptase and sequencing confirmed that at least the *CAC1* transcript was highly accumulated in *ddm1* (unpublished data). However, it remains to be seen whether transcriptional activation alone is sufficient to trigger the *ddm1*-induced transpositional activation of *CAC* elements; the *ddm1* mutation might also affect the epigenetic chromatin state, as the gene encodes a protein similar to the chromatin-remodelling factor SWI2/SNF2.

Maize transposons were the first plant genes found to be epigenetically regulated^{24,25}, and the epigenetic control of the maize *Spm* by *cis* elements and transposon-encoded Tnp proteins have been studied extensively²¹. Little is known, however, about controlling host factors. We report here endogenous *Arabidopsis* elements that are mobilized by a host gene mutation affecting epigenetic states. In addition to *ddm1*, a variety of *Arabidopsis* mutants in gene silencing and methylation^{26–32} provide excellent systems to dissect mechanistically the transpositional activation of *CAC* elements, which should further clarify the role of epigenetic controls on genome evolution.

Note added in proof: It has been shown recently that *Arabidopsis* elements similar to Robertson's *Mutator* in maize become active in the *ddm1* mutant³³. □

Methods

The growth conditions, preparation of plant genomic DNA and mapping of the *CLM* locus have been described¹². The suppression PCR method was as described¹⁶. We used *EcoRV*-digested genomic DNA for adapter ligation and subsequent amplification. The sequences of primers used for mapping the *CLM* locus, sequencing the *DWF4* gene, suppression PCR reactions, transposition monitoring (Fig. 2) and probe amplification (Fig. 3) are listed as Supplementary Information or on <http://www.nig.ac.jp/labs/AgriGen/cacta-Nature.html>. The Supplementary Information also includes sequences flanking the transposed *CAC* copies (Fig. 4).

Received 11 January; accepted 21 February 2001.

- SanMiguel, P. *et al.* Nested retrotransposon in the intergenic region of maize genome. *Science* **274**, 765–768 (1996).

- Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
- Martienssen, R. Transposons, DNA methylation and gene control. *Trends Genet.* **14**, 263–264 (1998).
- Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genet.* **20**, 116–117 (1998).
- Matzke, M. A., Mette, M. F., Aufsatz, W., Jakovitsch, J. & Matzke, A. J. Host defenses to parasitic sequences and the evolution of epigenetic control mechanisms. *Genetica* **107**, 271–287 (1999).
- Vongs, A., Kakutani, T., Martienssen, R. A. & Richards, E. J. *Arabidopsis thaliana* DNA methylation mutants. *Science* **260**, 1926–1928 (1993).
- Jeddeloh, J. A., Stokes, T. L. & Richards, E. J. Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nature Genet.* **22**, 94–97 (1999).
- Mittelsten-Scheid, O., Afsar, K. & Paszkowski, J. Release of epigenetic gene silencing by trans-acting mutations in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **95**, 632–637 (1998).
- Jeddeloh, J. A., Bender, J. & Richards, E. J. The DNA methylation locus DDM1 is required for maintenance of gene silencing in *Arabidopsis*. *Genes Dev.* **12**, 1714–1725 (1998).
- Steimer, A. *et al.* Endogenous targets of transcriptional gene silencing in *Arabidopsis*. *Plant Cell* **12**, 1165–1178 (2000).
- Hirochika, H., Okamoto, H. & Kakutani, T. Silencing of retrotransposons in *Arabidopsis* and reactivation by the *ddm1* mutation. *Plant Cell* **12**, 357–369 (2000).
- Kakutani, T., Jeddeloh, J., Flowers, S., Munakata, K. & Richards, E. Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proc. Natl Acad. Sci. USA* **93**, 12406–12411 (1996).
- Kakutani, T. Genetic characterization of late-flowering traits induced by DNA hypomethylation mutation in *Arabidopsis thaliana*. *Plant J.* **12**, 1447–1451 (1997).
- Soppe, W. *et al.* The gain of function epi-mutant FWA causes late flowering. *Mol. Cell* **6**, 791–802 (2000).
- Jacobsen, S., Sakai, H., Finnegan, E., Cao, X. & Meyerowitz, E. Ectopic hypermethylation of flower-specific genes in *Arabidopsis*. *Curr. Biol.* **10**, 179–186 (2000).
- Kakutani, T., Munakata, K., Richards, E. & Hirochika, H. Meiotically and mitotically stable inheritance of DNA hypomethylation induced by *ddm1* mutation of *Arabidopsis thaliana*. *Genetics* **151**, 831–838 (1999).
- Choe, S. *et al.* The *DWF4* gene of *Arabidopsis* encodes a cytochrome P450 that mediates multiple 22 α -hydroxylation steps in brassinosteroid biosynthesis. *Plant Cell* **10**, 231–243 (1998).
- Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. & Lukyanov, S. A. An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res.* **23**, 1087–1088 (1995).
- Nacken, W. K., Piotrowski, R., Saedler, H. & Sommer, H. The transposable element Tam1 from *Antirrhinum majus* shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol. Gen. Genet.* **228**, 201–208 (1991).
- Chandler, V. L. & Walbot, V. DNA modification of a maize transposable element correlates with loss of activity. *Proc. Natl Acad. Sci. USA* **83**, 1767–1771 (1986).
- Fedoroff, N. in *Epigenetic Mechanisms of Gene Regulation* (eds Riggs, A., Martienssen, R. & Russo, V.) 575–592 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1996).
- Scortecci, K. C., Dessaux, Y., Petit, A. & Van Sluys, M. A. Somatic excision of the Ac transposable element in transgenic *Arabidopsis thaliana* after 5-azacytidine treatment. *Plant Cell Physiol.* **38**, 336–343 (1997).
- Lawson, E. J., Scofield, S. R., Sjodin, C., Jones, J. D. & Dean, C. Modification of the 5' untranslated leader region of the maize Activator element leads to increased activity in *Arabidopsis*. *Mol. Gen. Genet.* **245**, 608–615 (1994).
- McClintock, B. Chromosome organization and genic expression. *Cold Spring Harbor Symp. Quant. Biol.* **16**, 13–47 (1951).
- McClintock, B. The suppressor-mutator system of control of gene action in maize. *Carnegie Institution of Washington Year Book* **57**, 415–429 (1958).
- Finnegan, E., Peacock, J. & Dennis, E. Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development. *Proc. Natl Acad. Sci. USA* **93**, 8449–8454 (1996).
- Ronemus, M., Galbiati, C., Ticknor, J., Chen, J. & Dellaporta, S. Demethylation-induced developmental pleiotropy in *Arabidopsis*. *Science* **273**, 654–657 (1996).
- Furner, I. J., Sheikh, M. A. & Collett, C. E. Gene silencing and homology-dependent gene silencing in *Arabidopsis*. Genetic modifiers and DNA methylation. *Genetics* **149**, 651–662 (1998).
- Amedeo, P., Habu, Y., Afsar, K., Scheid, O. M. & Paszkowski, J. Disruption of the plant gene MOM releases transcriptional silencing of methylated genes. *Nature* **405**, 203–206 (2000).
- Mourrain, P. *et al.* *Arabidopsis* SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* **101**, 533–542 (2000).
- Dalmay, T., Hamilton, A., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent RNA polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell* **101**, 543–553 (2000).
- Fagard, M. *et al.* AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl Acad. Sci. USA* **97**, 11650–11654 (2000).
- Singer, T., Yordan, C. & Martienssen, R. A. Robertson's Mutator transposons in *A. thaliana* are regulated by the chromatin-remodeling gene *Decrease in DNA methylation (DDM1)*. *Genes Dev.* **15**, 591–602 (2001).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank K. Munakata for technical assistance, A. Miyao for advice on the suppression PCR technique, and R. Martienssen and Y. Hiromi for comments on the manuscript. We acknowledge ABRC at the Ohio State University for the BAC clones.

Correspondence and requests for material should be addressed to T.K. (e-mail: tkakutan@lab.nig.ac.jp). The *CAC1*, *CAC2*, *CAC3* and *CAC4* sequences are deposited in GenBank under accession numbers AB052792, AB052793, AB052794 and AB052795, respectively.

The structural basis of Arfaptn-mediated cross-talk between Rac and Arf signalling pathways

C. Tarricone^{*†}, B. Xiao^{*}, N. Justin^{*}, P. A. Walker, K. Rittinger, S. J. Gamblin & S. J. Smerdon

Division of Protein Structure, National Institute for Medical Research, Mill Hill, London NW7 1AA, UK

^{*} These authors contributed equally to this work

Small G proteins are GTP-dependent molecular switches that regulate numerous cellular functions¹. They can be classified into homologous subfamilies that are broadly associated with specific biological processes. Cross-talk between small G-protein families has an important role in signalling, but the mechanism by which it occurs is poorly understood². The coordinated action of Arf and Rho family GTPases is required to regulate many cellular processes including lipid signalling³, cell motility⁴ and Golgi function⁵. Arfaptn is a ubiquitously expressed protein implicated in mediating cross-talk between Rac (a member of the Rho family) and Arf small GTPases. Here we show that Arfaptn binds specifically to GTP-bound Arf1 and Arf6, but binds to Rac-GTP and Rac-GDP with similar affinities. The X-ray structure of Arfaptn reveals an elongated, crescent-shaped dimer of three-helix coiled-coils. Structures of Arfaptn with Rac bound to either GDP or the slowly hydrolysable analogue GMPPNP show that the switch regions adopt similar conformations in both complexes. Our data highlight fundamental differences between the molecular mechanisms of Rho and Ras family signalling, and suggest a model of Arfaptn-mediated synergy between the Arf and Rho family signalling pathways.

Rac is central to cellular events that rely on plasma-membrane-based cell motility, and many of its downstream effectors have been identified. Arf family GTPases also seem to be required for several Rac-mediated functions, as exemplified by the cooperative activation of phospholipase D (PLD) activity by Arf1 and Rac⁶, and the requirement of Arf6 for Rac-mediated membrane ruffling⁷. Arfaptn (also known as partner of Rac (Por)) was identified as an effector of Rac⁹, and independently of Arf small GTPases¹⁰, and has been proposed as a potential mediator of cross-talk between Rac and Arf activities⁸. At present, three human isoforms are known (Arfaptn1/Por2, Arfaptn2/Por1 and Arfaptn1b), together with homologues from *Drosophila melanogaster* and *Caenorhabditis elegans* (Fig. 1a). Downstream targets of Arfaptn have not been identified; however, Arfaptn is an inhibitor of Arf1-dependent PLD activity¹¹, and the expression of truncated protein forms blocks the formation of membrane ruffles that otherwise result from the expression of constitutively active mutants of either Arf6 or Rac⁸. Given the role of Arfaptn in processes involving both Arf and Rho family activities, we undertook a structural and biochemical investigation of its complexes with these two small G proteins.

The structure of a fragment of Arfaptn2 (hereafter Arfaptn) encompassing the entire predicted coiled-coil region was solved by a three-wavelength multiwavelength anomalous dispersion (MAD) experiment (see Supplementary Information). The two Arfaptn monomers in the crystallographic asymmetric unit each consist of three α -helices (A–C) arranged as an extended antiparallel α -helical bundle (Fig. 1b). The monomers associate to form a highly elongated, crescent-shaped dimer, with an end–end distance of about 140 Å. Arfaptn monomers associate such that parts of helices

A, B and C from one monomer pack against helices A' and B' from the other. Although the helical packing interactions between subunits is largely hydrophobic in nature, the centre of the dimer contains a solvent-filled cavity. As far as we are aware, no structural homologues of the Arfaptn dimer have been reported.

Isothermal titration calorimetry (ITC) measurements (Fig. 2a; and Supplementary Information) reveal that Rac-GMPPNP and Rac-GDP bind to Arfaptn with dissociation constants (K_d) of 4.5 μ M and 3 μ M, respectively, and a stoichiometry of one G protein per Arfaptn dimer. The finding that Rac-GDP binds Arfaptn contradicts the original report of Rac binding to Arfaptn⁹ but confirms subsequent data¹². Our ITC data show that Arfaptn is not an effector of Rac in the generally accepted sense. Nevertheless, the interaction is specific for Rac (among the Rho family), as shown by ITC, because neither the GDP- nor GMPPNP-bound forms of Rho or Cdc42 bind detectably to Arfaptn. We therefore decided to determine the structures of both nucleotide-bound forms of Rac in complex with Arfaptn.

Initial efforts to crystallize Arfaptn in complex with wild-type Rac bound to the slowly hydrolysable GTP analogue, GMPPNP, resulted in two crystal forms, which we solved by molecular replacement. In both cases, initial electron density maps indicated loss of the γ -phosphate and a different switch I conformation to that expected for GMPPNP-bound Rac. This resulted from hydrolysis of the GMPPNP under the conditions used for crystallization, and these structures therefore contain GDP with a β -amidophosphate group (hereafter referred to as GDP). Both structures show that one molecule of Rac binds to each dimer of Arfaptn (Fig. 3a), as predicted from binding measurements. Rac sits roughly at the mid-point of the Arfaptn dimer and interacts largely through its switch I and II regions, burying $\sim 1,600$ Å² of solvent-accessible surface. The packing of switch II of Rac against four turns of helix A of Arfaptn monomer 'A' (residues 148–163) accounts for a large part of the overall interaction. The remaining interactions are mediated by switch I of Rac with a similar region of helix A. Details of the molecular interactions are shown in Fig. 3b.

In both crystal forms switch I is well ordered where it contacts Arfaptn, but residues 29–34 are poorly ordered in the tetragonal crystal form (the electron density shows the direction of the main chain, but does not define the side chains). This segment is well ordered in the monoclinic crystal form, however, because of crystal lattice contacts. The observation that switch I is ordered by contact with Arfaptn (and lattice contacts) is consistent with the innate mobility ascribed to this region in solution.

There is a single interaction between Rac and Arfaptn 'B' monomer, and inspection of the structure shows that binding of Rac to one of the two potential sites on the Arfaptn dimer prevents occupation of the other, hence the observed stoichiometry. Comparison of the sequences of Rac, Rho and Cdc42 suggests why Arfaptn is selective for Rac. Apparently, Rho does not bind because the arginine residue at position 5, which is not present in Rac, would clash with the A helix of Arfaptn. The sequences of Rac and Cdc42, in the switch I and II regions in particular, are very similar, but Ile 33 and Ser 41 of Rac are not conserved. Both of these residues are situated immediately adjacent to the Rac–Arfaptn interface, and it seems that these residues influence the local conformation of switch I and thereby facilitate the binding of Rac, and not Cdc42.

Several structures are now available for small G proteins in complex with binding domains derived from effector molecules. Rho family effectors can be loosely divided into those that are specific for Rho and those that bind preferentially to Rac/Cdc42. Many of the effectors for Rac/Cdc42 are characterized by the presence of CRIB (Cdc42/Rac interactive binding) domains that are crucial for PAK, ACK and WASP (p21-activated kinase, activated Cdc42-binding kinase and Wiskott–Aldrich syndrome protein, respectively) kinase autoregulation. CRIB domains are not ordered in isolation but instead fold either onto the surface of their

[†] Present address: European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy.

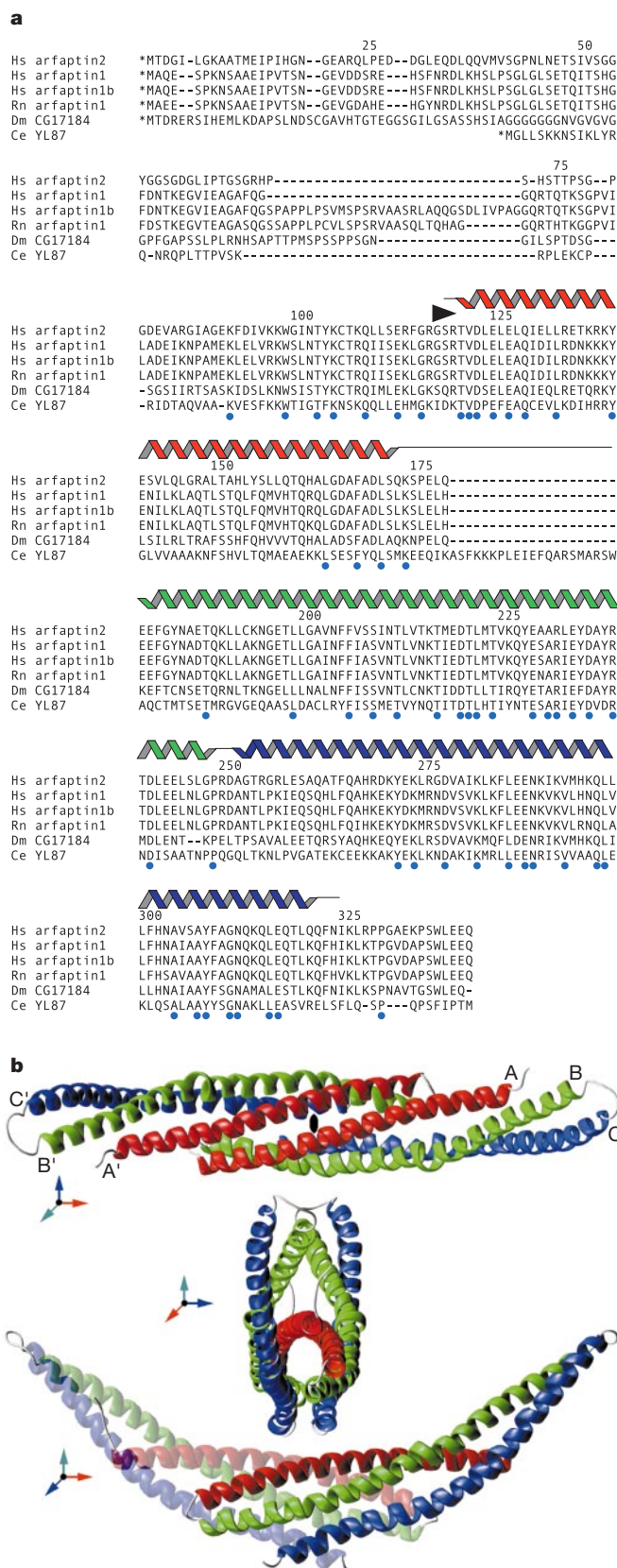


Figure 1 Sequence homology and structure of Arfaptin. **a**, Sequence homology between five Arfaptin homologues. The α -helical elements derived from the crystal structure are indicated and coloured as in **b**. The N terminus of the Arfaptin fragment used in this study, which encompasses the entire predicted coiled-coil region of these molecules, is indicated by the black triangle. Residues absolutely conserved between the six Arfaptin homologues are indicated by blue circles. **b**, Three orthogonal views of the Arfaptin dimer

in ribbons representation²⁶. Top, Arfaptin dimer viewed along its dyad axis. Helices A, B and C of each monomer are red, green and blue, respectively, and the dimer-related helices are labelled A', B' and C'. Middle, Arfaptin dimer viewed along the long axis, illustrating the cavity created by the five-helix barrel at the dimer interface. Bottom, Arfaptin viewed with its long axis horizontal and the dyad axis vertical, showing the crescent-like shape of the dimer.

parent molecule or their activating G protein. In contrast, many Rho-specific effectors have been predicted to bind to the GTPase through coiled-coil regions. Notably, the interaction between Arfapatin and Rac is quite unlike Rac-CRIB complexes and most closely resembles that of the Rho-binding domain of protein kinase N (PKN) complexed with RhoA¹³, in which the antiparallel coiled-

coil region of PKN packs against the helical segment of switch II.

To circumvent GMPPNP hydrolysis by wild-type Rac, we used the hydrolysis-impaired Q61L mutant of Rac to obtain crystals of the Rac-GMPPNP-Arfapatin complex. Unbiased electron density maps showed unequivocally that the γ -phosphate of the GMPPNP is present (Fig. 4a). The overall structure of Rac-GMPPNP-Arfapatin is very similar to that of Rac-GDP-Arfapatin and, in particular, there is little difference in either of the two switch regions (Fig. 4b). Given the similarity in the affinities of Rac-GDP and Rac-GMPPNP for Arfapatin, it follows that the molecular surface of Arfapatin can readily make the same interactions with Rac irrespective of whether GTP or GDP is bound. A key observation is that in the Arfapatin complexes the side chain of Thr 35 does not interact with nucleotide-associated Mg^{2+} , as it does in all other Rac-GTP containing structures, but instead interacts with Arfapatin. The orientation of Thr 35, and consequently the absence of interaction of its side chain with nucleotide-associated Mg^{2+} , is characteristic of GDP-bound GTPases. Modelling of the structure of Rac-GMPPNP, from the structure of its complex with the TPR domain of p67^{phox} (ref. 14), onto the G-protein component of the Rac-Arfapatin complex results in clashes of residues 36 and 37 from switch I of Rac with a segment of helix A (59–63) of Arfapatin. Thus, Arfapatin cannot accommodate a canonical GTP-bound conformation of Rac (Fig. 4b).

The similarity in the switch I conformations between the two different Rac-Arfapatin complexes leads us to propose that a related mechanism may be exploited by other proteins that bind, through contact with the switch regions, to both the GTP- and GDP-bound forms of small GTPases. For example, guanine-nucleotide dissociation inhibitors (GDIs) have crucial roles in small GTPase regulation by binding to the switch regions of both nucleotide-bound forms, thus directly inhibiting their signalling activities¹⁵. Crystal structures exist for the GDP-bound forms of Cdc42 (ref. 16) and Rac¹⁷

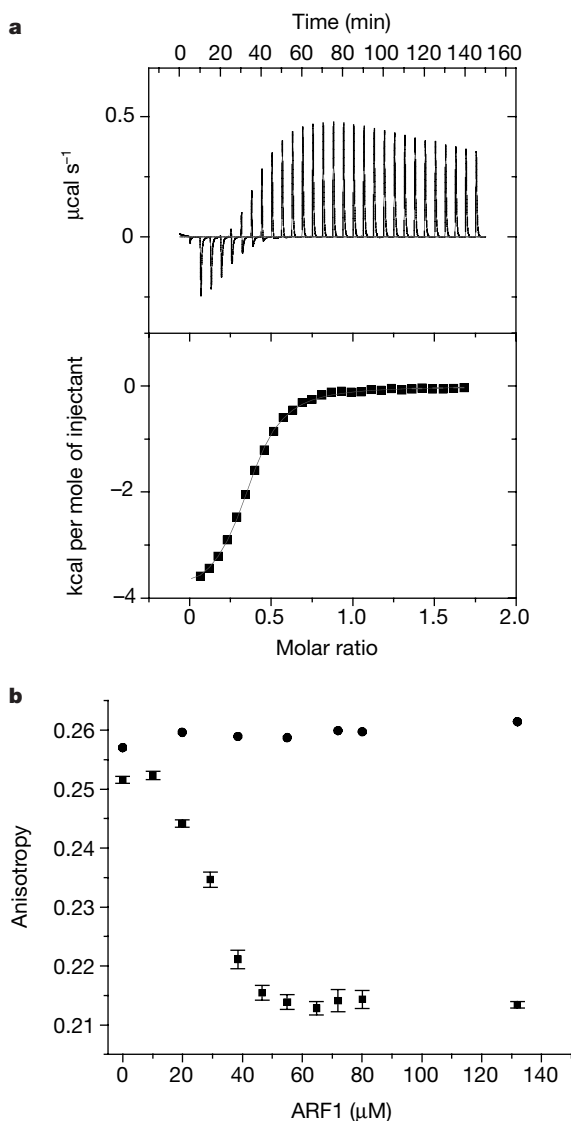


Figure 2 Arfapatin binding and competition studies. **a**, Isothermal titration calorimetry measurements of Rac1 binding to Arfapatin. Top, raw data of an ITC experiment performed at 22 °C. Bottom, integrated heat changes, corrected for the heat of dilution (the binding reaction and the heat of dilution have different signs), and the fitted curve based on a single-site model. For each type of experiment the following titrations were performed: GTPase into Arfapatin and Arfapatin into GTPase to confirm the stoichiometry of interaction. Figure shows the titration of Rac1-GDP at 1.0 mM into 0.06 mM Arfapatin dimer, which gives a K_d of 3 μ M and a binding ratio of 0.4 (Racs per Arfapatin monomer). Data for the titration of Arfapatin into Rac-GMPPNP are given in Supplementary Information. **b**, Arf1 competition for Rac binding to Arfapatin monitored by fluorescence anisotropy from Rac-mant-nucleotide. Figure shows titration of Arf1-GMPPNP (filled squares) and Arf1-GDP (filled circles) into Rac1-mant-GMPPNP-Arfapatin at 30 μ M. Error bars for the Arf1-GMPPNP titration represent the standard error of three measurements. The half-way point of the signal change corresponds roughly to the concentration of the mixture in the cuvette as would be expected for competing species with similar affinities for Arfapatin. Titrations using Rac-GDP (non-mant) also gave a reduction in anisotropy with similar start and end points. Data showing titration of Arf1 into Rac-mant-GDP-Arfapatin are given in Supplementary Information.

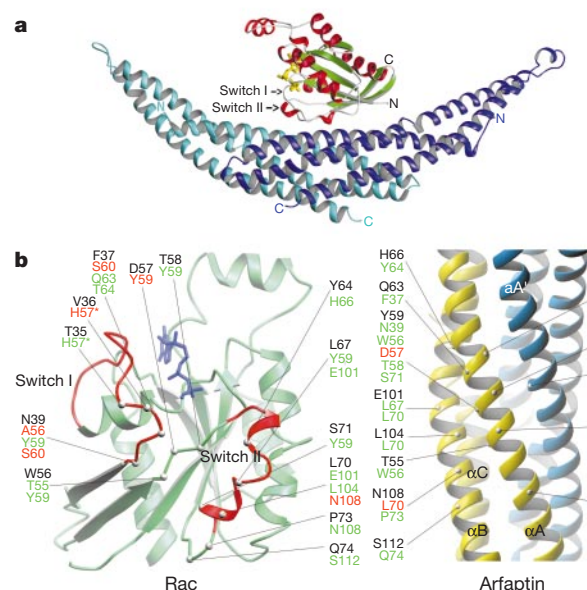


Figure 3 Structure of the Rac-Arfapatin complex. **a**, Rac-Arfapatin complex shown in ribbons representation, with Arfapatin in the same orientation as shown in Fig. 1b, bottom. Helices of the Rac are red, β -strands are green and the nucleotide is in yellow ball-and-stick representation. One monomer of the Arfapatin dimer is shown in blue, the other in pink. **b**, Arfapatin-Rac interface shown as an 'open book' representation. The $C\alpha$ positions of residues that interact are indicated by white spheres. Residue type and number are shown in black type with interacting residues from the other protein indicated in red (hydrogen-bonding interactions) or green (non-polar/van der Waals interactions). Asterisk denotes His 57 from Arfapatin molecule 'B' of the dimer (blue); all other Arfapatin residues are contributed from molecule 'A' (yellow). The two 'switch' regions of Rac are highlighted in red.

bound to their cognate GDIs. In both cases, the arrangement of the two proteins is such that the GDI components cannot accommodate the expected GTP-bound conformation of the GTPase. On the basis of these observations and our structural data, we predict that GDIs bind to GTP-bound Rho family GTPases by inducing a GDP-like conformation of switch I. This may represent a general mechanism by which certain target proteins can interact with the switch regions of both GDP- and GTP-bound forms of small GTPases.

Our structural results suggest that the behaviour of switch II of Rac is different to that of Ras—the prototypical small GTPase. Studies of Ras structure and function have largely defined the signalling paradigm for the whole superfamily¹⁸, in which effectors bind tighter to GTP-bound rather than GDP-bound GTPases. These differences in affinity have been ascribed to the conformational changes in the switch I and II regions on going from the GTP- to GDP-bound state, notwithstanding the fact that the switch regions of Ras are mobile in solution whether bound to GTP or GDP. The structure of the complex between Ras and phosphatidylinositol-3-OH kinase, together with supporting biochemical data¹⁹, show that switch II of Ras has an important signalling role in this system. However, there is a close structural similarity in switch II between the two Rac complexes presented here, uncomplexed Rac-GMPPNP, and the GDP- and GTP-bound forms of Cdc42 (ref. 16). Together, these structures imply that conformational changes in switch II have a less important role in signalling by Rho family small G proteins.

Because there is ambiguity in the studies concerning Arf6 interactions with Arfaptin, we determined the binding constants of the GDP and GTP forms of both Arf1 and Arf6 for Arfaptin by ITC. In contrast to Rac, only the GTP-bound forms of Arf1 and Arf6 bind to Arfaptin. Arf6 binding was found to be about tenfold weaker than Arf1 ($K_d \approx 2.5 \mu\text{M}$), in broad agreement with previous results¹⁰. Notably, our quantitative binding data establish that

Arf6 binds to Arfaptin *in vitro*, and that the binding is GTP dependent. These conclusions are an important corollary to previous data implicating Arf6 as an important binding partner of Arfaptin *in vivo*⁸.

Competition experiments, which monitored the fluorescence anisotropy of Rac-bound mant-nucleotide (Fig. 2b; and Supplementary Information), indicate that binding of Rac and Arf proteins to Arfaptin is mutually exclusive, indicating that they share at least part of the same binding site on the Arfaptin dimer. This observation suggests the mechanism by which Arfaptin directly mediates cross-talk between Arf and Rac GTPases. Arfaptin, which colocalizes with Rac, binds to and effectively sequesters Rac-GDP/GTP. In this sense Arfaptin behaves like a GDI, except that the Rac-binding activity of Arfaptin is, in turn, regulated by the local concentration of activated Arf. Activation of Arf leads to its binding to Arfaptin and displacement of Rac, thus releasing it for subsequent signalling activity. This scheme is particularly attractive in the light of experiments indicating the coupling of Rac signalling to Arf activation. For example, overexpressing EFA6, a guanine nucleotide exchange factor with high activity towards Arf6, promotes membrane ruffling and triggers the activation of Rac in spite of the fact that EFA6 is not, itself, an exchange factor for Rac²⁰. Moreover, it has been shown that Rac-induced membrane ruffling does not occur if Arf6 function is impaired and that this inhibition cannot be bypassed by forced recruitment of Rac to the plasma membrane⁴. Our model of Arfaptin function rationalizes these observations and provides a structural framework on which these ideas may be tested. □

Methods

Protein purification and crystallization

Wild-type Rac1, Rac1Δ8 (carboxy-terminal deletion), Rac1Q61L, RhoA Q63L, wild-type cdc42 and Arfaptin (118–341), together with full-length Arf1Q71L, Arf1I46D and Arf6Q67L, were prepared as glutathione S-transferase (GST) fusion proteins, cleaved from the GST overnight at 4 °C using either thrombin or PreScission protease, and further purified by gel-filtration (Superdex 75, Pharmacia). We expressed SeMet-labelled Arfaptin in *Escherichia coli* strain B834 (DE3).

Crystals of native and SeMet Arfaptin were obtained by vapour diffusion in hanging drops using a 1:1 mixture of protein ($\sim 10 \text{ mg ml}^{-1}$) and reservoir solution containing 0.1 M tri-sodium citrate dihydrate, 10% (v/v) iso-propanol and 50 mM sodium HEPES pH 7.5 at 4 °C. The SeMet MAD data were collected on ESRF BM14 and processed using DENZO & SCALEPACK²¹; from these data, SOLVE²² produced phases that were of sufficient quality to build a poly-alanine model. We used this model for molecular replacement calculations against the native data, and the resulting electron density map enabled a full model to be built using O²³. Refinement was carried out using REFMAC²⁴ and CNS²⁵. GTPase exchange with the slowly hydrolysable analogue of GTP, GMPPNP, was performed following standard methods and analysed by high performance liquid chromatography.

Complex crystals were grown by vapour diffusion at 18 °C using a 2:1 molar ratio of Arfaptin to Rac1; 0.5 ml of protein solution was mixed with 0.5 ml of reservoir buffer containing 0.1 M Tris pH 9.0 or 9.5 and 10% PEG 20K. Crystals were obtained in monoclinic and tetragonal space groups. Although both crystal forms grew from mixtures of Arfaptin and wild-type Rac that had been exchanged with GMPPNP, the rate of Rac hydrolysis of GMPPNP at the crystallization pH was sufficient to convert the nucleotide to GDP. To obtain the triphosphate nucleotide form of the Rac–Arfaptin complex, we used the hydrolysis impaired mutant Q61L of Rac. Tetragonal crystals of this complex that were similar to, but not isomorphous with, those described above were obtained under similar conditions.

Data for the three crystal types of the complex were collected at SRS and ESRF, and the structures were solved by molecular replacement using Amore²⁴ and refined using REFMAC and CNS. Extensive use was made of CNS simulated-annealing composite omit maps to avoid problems of phase bias. SDS–PAGE analysis of washed crystals showed proteolytic degradation of the Arfaptin component; this seems to be necessary for the formation of the tetragonal, but not the monoclinic form. Cleavage occurs at the amino-terminal end of helix B, remote from the site of Rac binding, and seems to allow the formation of a contact with a symmetry-related molecule.

Isothermal titration calorimetry

Binding of the GTPases to Arfaptin was measured by ITC using a MicroCal Omega VP-ITC isothermal titration calorimeter (MicroCal). GTPases at a concentration of 1.0 mM were titrated into 60 μM Arfaptin dimer. We repeated the experiments with 330 μM Arfaptin dimer titrated into 50 μM GTPase. All experiments were performed at 22 °C. The proteins were dialysed against 500 mM NaCl, 2 mM MgCl_2 , 0.5 mM β -mercaptoethanol and 50 mM Tris pH 8.7 (Arf1/Rac1) or 50 mM HEPES pH 7.5 (Arf6). After subtraction of the

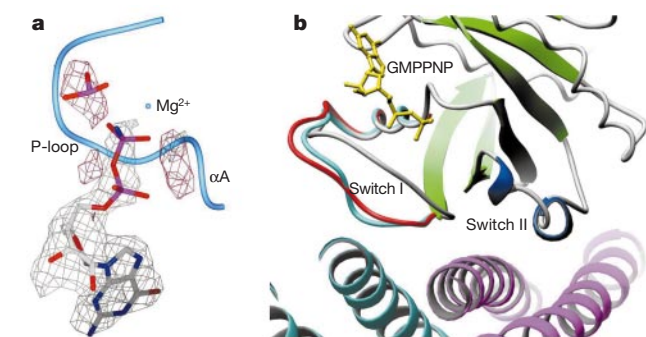


Figure 4 Switch I conformations of Rac in complexes with Arfaptin. **a**, Two types of electron density map around the guanine nucleotide moiety for the GTP- and GDP-bound forms of the Rac–Arfaptin complex, superimposed on a ball-and-stick representation of GDP. The P-loop and C-terminal parts of helix A of Rac, shown in a worm representation, are coloured cyan; the position of the Mg^{2+} ion in the Rac-GDP–Arfaptin complex is also indicated. The unbiased electron density ($2F_o - F_c$) for the GDP from the Rac-GDP–Arfaptin complex is shown in grey. Superimposed is an unbiased real-space difference electron density map (GMPPNP–GDP) in red, which clearly shows strong residual electron density at the expected location of the γ -phosphate. Additional electron density at the C terminus of helix A is due to a slight, but significant, movement of this helix between the Rac-GTP and Rac-GTP complexes of Arfaptin. **b**, Ribbon representation of part of the Arfaptin–Rac complex showing the interface of the two proteins viewed roughly in the same orientation as the Arfaptin in Fig. 1b, middle. Regions of the A and B helices of Arfaptin that contribute to the interface are shown in pink for one monomer and blue for the other. Rac is shown with the β -strands in green, switch II in blue and the guanine nucleotide in yellow ball-and-stick form. Three versions of switch I, shown in red, blue and grey, are taken from Rac-GDP–Arfaptin, Rac-GMPPNP–Arfaptin and Rac-GTP/p67^{phox} (ref. 14) structures, respectively. The parts of Rac not included in switch I are essentially the same in these three crystal structures and only one of them has been shown for clarity.

dilution heats, calorimetric data was analysed using the evaluation software MicroCal Origin v5.0 (MicroCal Software). Negative binding results from ITC (for Arf-GDP, Rho-GDP/GTP and Cdc42-GDP/GTP) were confirmed by either fluorescence anisotropy (using mant-nucleotide), surface plasmon resonance, or both.

Fluorescence anisotropy titrations

Competition experiments were performed using an ISS PC1 spectrofluorimeter. All measurements were taken at 20 °C with $\lambda_{\text{exc}} = 366$ nm. We viewed emission through a KV399 cutoff filter, and recorded the measurements using the L format of the instrument. Rac1 was loaded with either mant-GDP or mant-GMPPNP, and mixed with stoichiometric amounts of Arap1 (dimer). The mixture was placed in the cuvette and titrated against either Arf1-GDP, Arf1-GMPPNP, Rac1-GDP or Rac1-GMPPNP while the change in fluorescence anisotropy was monitored.

Received 1 November 2000; accepted 5 March 2001.

- Macara, I. G., Lounsbury, K. M., Richards, S. A., McKiernan, C. & Bar-Sagi, D. The Ras superfamily of GTPases. *FASEB J.* **10**, 625–630 (1996).
- Bar-Sagi, D. & Hall, A. Ras and Rho GTPases: A family reunion. *Cell* **103**, 227–238 (2000).
- Fensome, A., Whatmore, J., Morgan, C., Jones, D. & Cockcroft, S. ADP-ribosylation factor and Rho proteins mediate fMLP-dependent activation of phospholipase D in human neutrophils. *J. Biol. Chem.* **273**, 13157–13164 (1998).
- Zhang, Q., Calafat, J., Janssen, H. & Greenberg, S. ARF6 is required for growth factor- and rac-mediated membrane ruffling in macrophages at a stage distal to rac membrane targeting. *Mol. Cell. Biol.* **19**, 8158–8168 (1999).
- Wu, W. J., Erickson, J. W., Lin, R. & Cerione, R. A. The γ -subunit of the coatamer complex binds Cdc42 to mediate transformation. *Nature* **405**, 800–804 (2000).
- Walker, S. J., Wu, W. J., Cerione, R. A. & Brown, H. A. Activation of phospholipase D1 by Cdc42 requires the Rho insert region. *J. Biol. Chem.* **275**, 15665–15668 (2000).
- Boshans, R. L., Szanto, S., van Aelst, L. & D'Souza-Schorey, C. ADP-ribosylation factor 6 regulates actin cytoskeleton remodeling in coordination with Rac1 and RhoA. *Mol. Cell. Biol.* **20**, 3685–3694 (2000).
- D'Souza-Schorey, C., Boshans, R. L., McDonough, M., Stahl, P. D. & Van Aelst, L. A role for POR1, a Rac1-interacting protein, in ARF6-mediated cytoskeletal rearrangements. *EMBO J.* **16**, 5445–5454 (1997).
- Van Aelst, L., Joneson, T. & Bar-Sagi, D. Identification of a novel Rac1-interacting protein involved in membrane ruffling. *EMBO J.* **15**, 3778–3786 (1996).
- Kanoh, H., Williger, B. T. & Exton, J. H. Arfap1, a putative cytosolic target protein of ADP-ribosylation factor, is recruited to Golgi membranes. *J. Biol. Chem.* **272**, 5421–5429 (1997).
- Williger, B. T., Ostermann, J. & Exton, J. H. Arfap1, an ARF-binding protein, inhibits phospholipase D and endoplasmic reticulum/Golgi protein transport. *FEBS Lett.* **443**, 197–200 (1999).
- Joneson, T., McDonough, M., Bar-Sagi, D. & Van Aelst, L. RAC regulation of actin polymerization and proliferation by a pathway distinct from Jun kinase. *Science* **274**, 1374–1376 (1996).

- Maesaki, R. *et al.* The structural basis of Rho effector recognition revealed by the crystal structure of human RhoA complexed with the effector domain of PKN/PRK1. *Mol. Cell* **4**, 793–803 (1999).
- Lapouge, K. *et al.* Structure of the TPR domain of p67^{phox} in complex with Rac-GTP. *Mol. Cell* **6**, 1–20 (2000).
- Geyer, M. & Wittinghofer, A. GEFs, GAPs, GDIs and effectors: taking a closer (3D) look at the regulation of Ras-related GTP-binding proteins. *Curr. Opin. Struct. Biol.* **7**, 786–792 (1997).
- Hoffman, G. R., Nassar, N. & Cerione, R. A. Structure of the Rho family GTP-binding protein Cdc42 in complex with the multifunctional regulator RhoGDI. *Cell* **100**, 345–356 (2000).
- Scheffzek, K., Stephan, I., Jensen, O. N., Illenberger, D. & Gierschik, P. The Rac–RhoGDI complex and the structural basis for the regulation of Rho proteins by RhoGDI. *Nature Struct. Biol.* **7**, 122–126 (2000).
- Wittinghofer, A. Signal transduction via Ras. *Biol. Chem.* **379**, 933–937 (1998).
- Pacloud, M. E. *et al.* Crystal structure and functional analysis of ras binding to its effector phosphoinositide 3-kinase γ . *Cell* **103**, 931–943 (2000).
- Franco, M. *et al.* EFA6, a sec7 domain-containing exchange factor for ARF6, coordinates membrane recycling and actin cytoskeleton organization. *EMBO J.* **18**, 1480–1491 (1999).
- Otwinski, Z., Minor, W. in *Data Collection and Processing* (eds Sawyer, L., Isaacs, N. & Bailey, S.) 556–562 (Daresbury Laboratory, Warrington, UK, 1993).
- Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
- Jones, T. A., Zhou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Collaborative Computer Project No. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Kraulis, P. J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946–950 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We acknowledge S. Price for assistance with protein purification; G. Dodson and A. Lane for critically reading the manuscript; T. Magee for helpful discussions; D. Jones for expression clones of Arf; the staff of SRS Daresbury and ESRF Grenoble, particularly G. Leonard, for beamtime allocation and assistance with X-ray data collection; and J. F. Eccleston for assistance with fluorescence experiments.

Correspondence and requests for materials should be addressed to S.J.G. (e-mail: sgambli@nimr.mrc.ac.uk). Coordinates have been deposited in the Brookhaven Database under accession codes 1I49 (Arfap1), 1I4D (Arfap1–RacGDP monodimer form), 1I4L (Arfap1–RacGDP tetragonal form) and 1I4T (Arfap1–RacGMPPNP).

errata

Drought-induced guard cell signal transduction involves sphingosine-1-phosphate

Carl K.-Y. Ng, Kathryn Carr, Martin R. McAinsh, Brian Powell & Alistair M. Hetherington

Nature **410**, 596–599 (2001).

In the second paragraph of this paper, (*trans*-4-sphingenine, d18:1 (ref. 4)) should have read (*trans*-4-sphingenine, d18:1⁴); (d18:1 (ref. 8)) should have read (d18:1⁸); (d18:2 (refs 4,8)) should have read (d18:2^{4,8}); and (t18:1 (ref. 10)) should have read (t18:1⁸)¹⁰. □

BRI1 is a critical component of a plasma-membrane receptor for plant steroids

Zhi-Yong Wang, Hideharu Seto, Shozo Fujioka, Shigeo Yoshida & Joanne Chory

Nature **410**, 380–383 (2001).

In Figure 3b, the label for CIP at the third lane should be + rather than –. □

New cart on the block

A selection of equipment for general laboratory use.

Prima UBD and UBR

Grant Instruments www.grant.co.uk

Dry block heaters for everyday use

Prima microprocessor-controlled laboratory heaters are available in six models suitable for one, two or four blocks in two series. UBR models offer a data monitoring capability via an RS 232 interface and a built-in timer with programmable delayed start. The heaters can accommodate samples in tubes, vials or slides for temperature control over a range from just above ambient to 130°C. Heating is rapid, rising from ambient to 100°C in 16 minutes. An optional external probe is available for precise in-block temperature control and an auto two-point calibration facility is also provided. All units have an angled display panel and are housed in a waterproof enclosure.

Precision-Aire

Fabreeka www.fabreeka.com

No motions on this table

Designed to hold sensitive lab equipment, Precision-Aire isolation tables provide a self-levelling work surface and low-frequency isolation in areas where footfall and structural vibration can interfere with reliability and accuracy. They can support microscopes, electro-optical and vision systems, micro-positioning and wafer inspection, and are available in a range of sizes to accommodate loads of 170 to 1,100 kg. Work surface options include stainless steel, plastic laminate and breadboards with mounting holes. Accessories include shelves, monitor arms, lean bars, castors and a variety of enclosures.

Glassware Cart

Labconco www.labconco.com

Ready to roll

Glassware, plastic ware and small instruments can be transported on Labconco's latest cart, which is equipped with either two large or four small vinyl-coated wire baskets and a removable plastic drip pan to contain spills. The welded frame of 1-inch square tubing is coated with baked-on epoxy powder to resist corrosion and abrasion. Rubber casters assist steering and absorb vibration. At 19 inches wide, the cart fits through most doors and narrow corridors, and is delivered ready-to-run.

HC-3000 HybriCycler

UVP www.uvp.com

The shape of things to come

The HybriCycler hybridization oven features a wrap-around design to allow complete visibil-



Block booking: Prima heaters from Grant Instruments of Cambridge.

ity of bottles or tubes. Shielding from β radiation is provided by sealed acrylic construction. The unit is compact, with multiple bottle size capacity and a removable stainless steel spill tray. It is designed for precision control of temperature at ambient to 80°C, with speed control at 11 r.p.m. The oven is recommended for all cDNA library screenings, primer synthesis and nucleic acid hybridization applications.

Unico 1200

Unico www.unico1.com

Upgraded spectrophotometer

An enhanced version of the the Unico Model 1100, the 1200 has a bandwidth of 5nm, a wavelength range of 325–1,000 nm, temperature/absorbance/concentration modes, and an automatic zeroing and blanking button. In addition, there is optional Windows-based software and external serial printing, tool-free

and alignment-free lamp change, and a large sample compartment for accessories. These include optional 10 mm single-cell or round test tube holder, V-type round test tube holder, water-jacketed flow-cell holder and panel and longpath cell holder for up to 50-mm-longcell. The Model 1200 is suitable for routine laboratory use.

751 GDP Titrino

Metrohm www.metrohm.co.uk

Titration à la carte

The screen on this titrator displays live titration curves as they are generated. As well as dynamic and monotonic equivalence point titrations, the unit can perform titrations to two pre-set end-points. Karl Fischer titrations, and pH stat applications are also supported. Titration methods can be saved onto a memory card that can be shared with remote labs to ensure consistent methodology. The titrator is supplied with a read-only memory card containing titration application methods, and a titration procedure (TIP) mode enables the user to select and link up to 30 instructions to form a customized procedure. The Titrino is able to control two 685 Dosimat dispensing units for titration and two 725 or 665 Dosimats for auxiliary dosing of solutions

Optima

Boekel Scientific www.boekelsci.com

Configurable circulating baths

Optima Series modular immersion circulating baths allow users to customize the bath's configuration by switching components to vary tank size, temperature range, accessories, and set-point precision. The series is built



See the heat: UVP's new HybriCycler.

new on the market

around five different circulator models. Temperature range is from -50°C (with companion immersion cooler) to $+200^{\circ}\text{C}$ (in a covered tank). Eight different steel or plastic tanks and accessories including lids, racks, cooling systems and insulating spheres are available. The base model offers analogue temperature control and no circulating pump. Other models are digitally controlled, with temperature precision to $\pm 0.02^{\circ}\text{C}$.

Research pipette system

Eppendorf www.eppendorf.com
Uniform advantage

The 'Research Family' of pipettes consists of adjustable, fixed-volume and multi-channel models, plus the electronic Research Pro. Volumes range from $0.1\ \mu\text{l}$ to 5 ml. Advantages claimed for the system include improved ergonomics, optimal pipetting of various volumes and uniform standards across different pipettes. Maintenance and service parts are interchangeable, and storage can be provided by a pipette carousel or holder that will accommodate all Research pipettes.

12-arm water maze

Columbus Instruments www.colinst.com
Rat race

This watertight maze is designed for behavioural studies with rats. Animals can be tracked automatically by video (Videomex-One or Videomex-V) as they swim through the maze. Constructed of black-painted stainless steel, the central area of the maze measures 685 mm, with arm length and width 380 and 130 mm, respectively, and a depth of 305 mm.

BE-1

Airfiltronix www.airfiltronix.com
High-visibility particle/fume hood

The Airfiltronix Balance Enclosure is a workstation for weighing, measuring, pouring, mixing and performing any process operation that involves dust, particulates, powders or vapours. The HEPA filter removes 99.99% of



particulates larger than 0.3 microns. Optional charcoal filters are available for adsorbing vapours. The workspace area measures 28 inches wide by 21 inches deep by 19 inches high and weighs approximately 80 pounds. The unit offers high visibility and adjustable face velocity.

Tube furnace

Carbolite www.carbolite.co.uk
Keep your powder hot

Carbolite manufactures an extensive range of tube furnaces including split, vertical, horizontal and multi-zone versions. The latest model is designed specifically for laboratory use, enabling powders to be continuously processed at up to $1,200^{\circ}\text{C}$ in an inert atmosphere. It takes up to 5 litres of material at a time in a rotating work tube that runs at 2–12 r.p.m. at zero to 10° from horizontal. Total length of the work tube is 1,800 mm.

LS 45

PerkinElmer www.perkinelmer.com
Luminescence 'spec' for routine analysis

The LS45 fixed-slit, monochromator-based spectrometer uses a high-energy, pulsed xenon source for excitation, minimizing photobleaching of samples. Accessories include a remote fibre optic device for non-destructive testing of fluorescent papers and fabrics and

hazardous materials. The AS-93plus autosampler provides a high-sensitivity alternative to microplate reading. The spectrometer is controlled using FL WinLab software. It is suitable for laboratories using fluorescence, phosphorescence, chemi- or bioluminescence techniques in clinical, environmental, pharmaceutical and industrial applications.

Durafuge 100

Precision www.precisionsci.com
Quiet operator

Generating only 57dBA at maximum speed, this small centrifuge, measuring 15 inches by 18 inches, has a tube capacity of 12 blood tubes (5, 7 or 10 ml) in a sealed, swing-out rotor. Power is provided by a maintenance-free induction motor. Safety features include sealed accessories certified for aerosol containment, a digital imbalance detection system and a double-lid interlock design.

SPACE

Envair www.envair.co.uk
Clean air and containment

SPACE (for Special Procedure and Containment Environment) is Envair's new double-sided Class II type safety cabinet with full unidirectional clean air in the work zone. The work surface of SPACE has recesses for work arriving in trays, or alternatively the recesses may be covered to create a flat working surface. SPACE is fitted a Vari-Height stand. When raised, this allows work to be loaded and unloaded on both sides of the cabinet with operators standing. When lowered, more intricate scientific procedures can be carried out on the 'knee-room' side of the cabinet with the operator seated in an ergonomically optimized position. SPACE has been designed to be decontaminated with formaldehyde or vaporized hydrogen peroxide, for which it has all the necessary sealing panels and connections.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENTS

**DIVERSE
SMALL
MOLECULES
READY FOR
SCREENING**

**10K–50K
Compounds
in Plates**

- Next Day Delivery*
- Competitively Priced

*(Some Restrictions Apply)

Chembridge Corporation • 800.964.6143
858.451.7400 • 858.451.7401 FAX • E-mail: chem@chembridge.com

www.chembridge.com

**BACs • YACs • PACs
cDNAs**

We provide clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at
www.resgen.com

CUSTOMISED

immunoassays for
research and production

BioGenes
Gesellschaft für Biopolymere mbH

Internet www.biogenes.de
Berlin *Germany
Phone +49 (0)30-65 76 23 96

Contacts

Publisher: Fabien Savenay
Editor: Paul Smaglik
Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:

Leonie Welss (4954)

European Manager:

Nevin Bayoumi (4978)

European Sales Consultant:

Jonathan Bull (4973)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Sille Opstrup (4994)

Production Manager:

Billie Franklin
To send films and materials use
London address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau

Tel + 33 (0) 1 43 20 16 51

Fax + 33 (0) 1 43 20 51 52

e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Amanda Wren

Tel + 34 91 3913835

Fax + 34 91 319 0091

e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner

Tel + 49 89 54 90 57 11/2

Fax + 49 89 54 90 57 20

e-mails: p.phelan@nature.com

k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: classified@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi Shinjuku-Ku,
Tokyo 162

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.cowan@nature.com

Japan Manager:

Kate Cowan

Women physicists needed

The representation of women in physics can be viewed as either a 'glass half-full' or 'glass half-empty' situation. The optimist would point out that the number of women faculty in physics has doubled in the past 20 years or so. The pessimist would note that, even so, women still hold less than 10% of physics faculty positions worldwide and their numbers in administration are even more minuscule.

If representation is poor on the university level, it is abysmal at the leadership levels of international physics societies, Judy Franz, executive officer of the American Physical Society (APS), said last month at the APS Washington meeting. For example, Franz notes that, until fairly recently, the International Union of Pure and Applied Physics (IUPAP) was dominated by men. Perhaps an unintended result of this make-up has been a failure to invite women speakers to international physics conferences. Without such role models, young women physicists attending conferences receive the subtle signal that there is little or no room for them at the end of the pipeline.

The other signals that cause women to leave the physics pipeline at a disproportionately higher rate than those following paths towards faculty positions in other disciplines, such as biology or social sciences, are poorly understood. But Franz doubts that it is because women are smarter than their male colleagues at realizing that more lucrative positions await them in sectors outside academia, or that women don't need physics PhDs to land them. "That doesn't ring true to me," Franz said.

The IUPAP has started a working group on women in physics to gather more data about conditions and trends in different countries. The group will discuss its findings at its first international meeting next March in Paris. Women, most assuredly, will be invited.

Paul Smaglik

Naturejobs editor



Contents

SPECIAL REPORT

America beckons as CERN enters a high-energy hiatus p4

FOCUS

Genetics p6

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific
job market

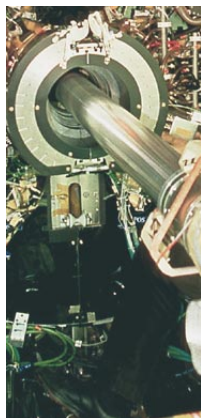
FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS



The loss of one of CERN's most powerful tools could change the dynamics of employment for high-energy physicists in Europe, says Alexander Hellemans.

The dismantling of the Large Electron–Positron Collider (LEP), which began earlier this year, left its users with mixed feelings. Based at CERN, the European Laboratory for Particle Physics near Geneva, LEP leaves a positive legacy because it helped physicists flesh out the standard model, the theoretical characterization of fundamental particles and their interactions. But shutting down LEP without quite capturing a glimpse of the model's missing piece, the elusive Higgs boson, left many wallowing in disappointment — and searching for new opportunities.

“A lot of people have invested something like 15 years of their professional life into either the accelerator or the experiments, and they feel quite deflated,” says Roger Bailey, who heads the operations group at CERN. Many believe that CERN has missed a major opportunity in refusing to extend the life of LEP by another 6 to 12 months, after physicists claimed to have found traces of the Higgs boson.

Decommissioning LEP has meant more than melancholy. It also ushered in some career questions. The more powerful Large Hadron Collider (LHC) will occupy the same 27-kilometre tunnel straddling the French–Swiss border, but how will the six-year gap between switching off LEP and starting up the LHC affect opportunities for high-energy physics jobs in Europe? Will more European high-energy physicists migrate to the United States?

Fermilab, near Chicago, may provide a draw for some (see *Nature* **409**, 754–755; 2001). Its Tevatron, today's most powerful proton–antiproton collider, will allow scientists who had to stop looking for the Higgs

boson at CERN to resume their quest without waiting for the LHC to come on line.

THE PHYSICS GAP

CERN now faces a relatively quiet period of five to six years during which the LHC will be built. And because LEP accounted for about a quarter of all the physics work at CERN, there is a large gap in opportunities for physicists. “Filling the gap between now and 2006 is clearly a bit of a problem,” says David Plane, spokesman for OPAL, a LEP experiment. But LEP's legacy will continue to occupy physicists. “There is still a lot of work to be done on the LEP data, but this will not completely fill the gap,” Plane says.

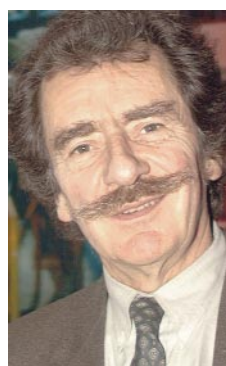
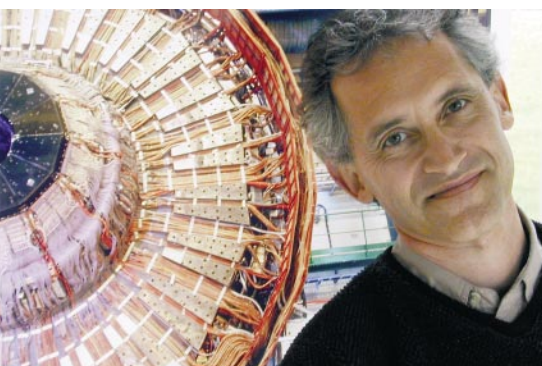
The construction and planning of the LHC and its detectors will involve many physicists. Yet more are designing simulation programs that will help prepare future LHC researchers for what to expect from the machine. Dieter Schlatter, spokesman for ALEPH, another LEP experiment, says there's still plenty of work to be done. “The construction of the detectors is such an enormous enterprise that there are not really enough people — we have to do experiments and invest in the future,” he says.

CERN had already been following a general reduction of staff, from about 3,500 in 1995 to 2,100 in 2005. About 2,500 now work at the site. But the 200 or so physics staff positions at CERN will only be affected to a minor degree, says Claude Detraz, who is CERN's director of research for all non-LHC experiments. Keeping up fundamental research and trying to maintain high-level science will be difficult

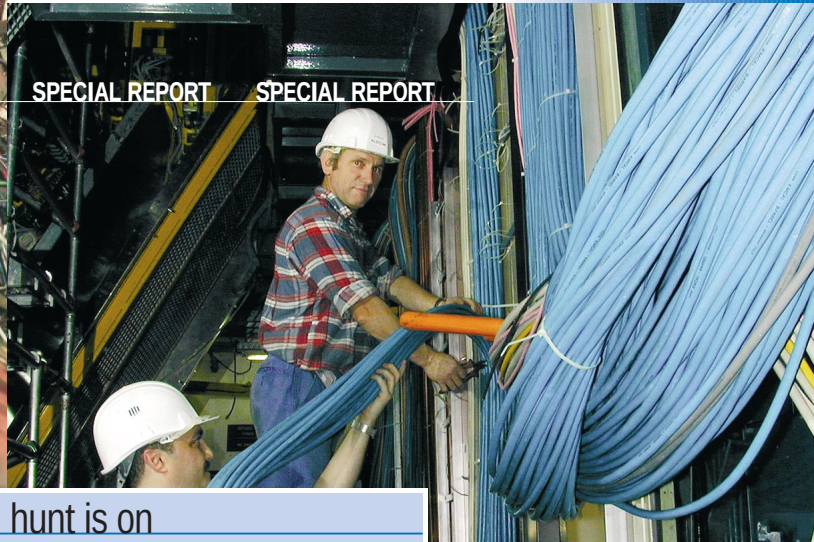
for CERN without one of its most powerful tools, says Detraz. CERN's directorate has been anticipating this problem for several years, and has been issuing an annual interim research plan.

The smaller experiments continuing at CERN cannot accommodate huge numbers of PhD candidates, but they do provide better experience than can be gained from the bigger experiments. The average PhD student doesn't even get close to the apparatus with bigger

In the hiatus before the Large Hadron Collider is ready, there's plenty of work to be done at CERN, say Roger Bailey (below), Dieter Schlatter (below centre) and Claude Detraz.



CERN



The hunt is on

Fermilab, near Chicago, is proving to be a big draw for high-energy physicists. After a major refit of one its detectors, DZero, the Tevatron has now been switched back on. In March, it officially started 'Run II' in a concerted effort to find the Higgs boson. An upgrading programme will intensify the beams, and

researchers will search for signs of the Higgs in the debris left by proton-antiproton collisions. The laboratory may be able to spot the elusive particle by collecting and integrating data over the next four or five years, perhaps beating the LHC by one or two years.
<http://www.fnal.gov>

Out with the old: CERN's Large Electron-Positron Collider gets dismantled.



Sau Lan Wu says young people need to deal with experimental data.



More interest: Al Goshaw has seen a rise in the numbers of applications for work at Fermilab.



Researcher mobility has been one of the characteristics of high-energy physics — detectors and collaborations are huge, attracting people from all over the world. CERN has been the strongest international attraction point over the past few decades. But the loss of LEP could cause a shift. "The flow is now more from Europe to the United States," says John Womersley, a DZero spokesman. "But I expect that this flow will be in the other direction 5 to 10 years from now."

However, there is a fresh dimension to this physics mobility. European institutions and national programmes are becoming more directly involved in the large US physics facilities. That development is forcing large US labs to become "truly international operations, something that SLAC and Fermilab have not really been in the past," says Womersley.

One of the reasons for this increased internationalization of science is that large facilities are now so expensive that duplication of machines in Europe and the United States is no longer practical. The migration of scientists between large machines will become commonplace. As the goals for particle physicists continue to shift, so too will the centres of employment.

Alexander Hellemans is a science writer in Naples.

<http://www.cern.ch>

<http://www.slac.stanford.edu>

CERN

FERMILAB

experiments because everything has been designed and built by engineers and technicians, says Rolf Landua, principal investigator of ATHENA, a project attempting to create antihydrogen atoms. "For the young physicists who want to 'rub' with physics, we have planned shorter experiments that allow them to get hands-on experience," agrees Detraz. And this is why Plane also encourages young physicists to get involved with the LHC at an early stage. "They should get detector hardware experience," Plane says. "I think that is all part of being a high-energy physicist."

HUNTING FOR ACTION

The LHC construction phase will not affect physicists who are firmly established in the field. A hiatus of six years for construction is relatively small compared with the time it can take to prepare and execute a large experiment, which can span 10 to 15 years, says Detraz. "I know physicists who have already been working for 11 years on LHC projects," he adds.

But the construction phase will affect the possibilities for young physicists. Many are looking for opportunities to work with experimental data elsewhere, especially at the Stanford Linear Accelerator (SLAC) and Fermilab. Young people need to deal with experimental data as well as with the hardware, says Sau Lan Wu of the University of Wisconsin, who leads a research group at SLAC's current flagship experiment, BaBar. "My students will get hardware training with LHC experiments, and will write their thesis using BaBar," she says.

The search for the Higgs boson is dictating where the action will be, and the near discovery of this particle at CERN has whetted the appetite of researchers. Many ex-OPAL people from CERN are now attached to Fermilab experiments such as DZero, a Tevatron collaboration that will look for signs of the Higgs. "There is a shift — the people want to go to laboratories working with the Fermilab experiment, which is where a lot of the action will be," says Plane.

Al Goshaw of Duke University in Gaithersburg, Maryland, spokesman for the CDF collaboration which will also look for the Higgs with the Tevatron at Fermilab, confirms an increase in applications. "We have received appreciable interest from people who made big contributions to one of the four LEP experiments," he says.

SLAC

FERMILAB